Four dimensional variational inversion of atmospheric chemical sources in WRFDA

by

J. J. Guerrette

B.S., Rochester Institute of Technology, 2009

M.S., Rochester Institute of Technology, 2011

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Mechanical Engineering

2016

This thesis entitled:
Four dimensional variational inversion of atmospheric chemical sources in WRFDA
written by J. J. Guerrette
has been approved for the Department of Mechanical Engineering

Daven K. Henze

Prof. Jana Milford

Asst. Prof. Peter Hamlington

Prof. Jose-Luis Jimenez

Dr. David Baker

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Guerrette, J. J. (Ph.D., Mechanical Engineering)

Four dimensional variational inversion of atmospheric chemical sources in WRFDA

Thesis directed by Assoc. Prof. Daven K. Henze

Atmospheric aerosols are known to affect health, weather, and climate, but their impacts on regional scales are uncertain due to heterogeneous source, transport, and transformation mechanisms. The Weather Research and Forecasting model with chemistry (WRF-Chem) can account for aerosol-meteorology feedbacks as it simultaneously integrates equations of dynamical and chemical processes. Here we develop and apply incremental four dimensional variational (4D-Var) data assimilation (DA) capabilities in WRF-Chem to constrain chemical emissions (WRFDA-Chem). We develop adjoint (ADM) and tangent linear (TLM) model descriptions of boundary layer mixing, emission, aging, dry deposition, and advection of black carbon (BC) aerosol. ADM and TLM model performance is verified against finite difference derivative approximations. A second order checkpointing scheme is used to reduce memory costs and enable simulations longer than six hours. We apply WRFDA-Chem to constraining anthropogenic and biomass burning sources of BC throughout California during the 2008 Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS) field campaign. Manual corrections to the prior emissions and subsequent inverse modeling reduce the spread in total emitted BC mass between two biomass burning inventories from a factor of $\times 10$ to only $\times 2$ across three days of measurements. We quantify posterior emission variance using an eigendecomposition of the cost function Hessian matrix. We also address the limited scalability of 4D-Var, which traditionally uses a sequential optimization algorithm (e.g., conjugate gradient) to approximate these Hessian eigenmodes. The Randomized Incremental Optimal Technique (RIOT) uses an ensemble of TLM and ADM instances to perform a Hessian singular value decomposition. While RIOT requires more ensemble members than Lanczos requires iterations to converge to a comparable posterior control vector, the wall-time of RIOT is $\times 10$ shorter since the ensemble is executed in parallel. This work demonstrates that RIOT im-

proves the scalability of 4D-Var for high-dimensional nonlinear problems. Overall, WRFDA-Chem and RIOT provide a framework for air quality forecasting, campaign planning, and emissions constraint that can be used to refine our understanding of the interplay between atmospheric chemistry, meteorology, climate, and human health.

Dedication

To my mother, Janine, for teaching me to persevere.

Acknowledgements

Thank you to my advisor, Daven Henze, who spent countless hours providing critiques, mentorship, and support during the most difficult parts of this work. He has helped lay the foundation for my career ahead. Thank you to Nicolas Bousserez for his consistent mathematical prowess and for including me in his work. Matt and Yanko made this work possible by keeping the group cluster running efficiently. Gregory Carmichael's contributions to this field and vision were integral to this project getting off the ground. Thank you Pablo Saide for providing insight and sharing your work. Scott Spak, Li Zhang, and Kateryna Lapina provided research and career advice. The whole research group has contributed in one way or another to my progress.

I am especially thankful for grants from the U.S. EPA's STAR grant R835037 and NOAA's FIREX campaign, which funded this research. This work does not necessarily reflect the views of those agencies, and no official endorsement should be inferred. Resources supporting this work were provided by the NASA HEC Program through the NAS Division at Ames Research Center. Thank you to NCAR and NOAA for maintaining WRF, WRF-Chem, and WRFDA. Thank you to NASA for supporting the ARCTAS mission, and Y. Kondo for providing black carbon observations. I used FIRMS data from the LANCE system operated by NASA/GSFC ESDIS, funded by NASA/HQ.

Thank you to my fellow PhD students, who have acted as a support network and mental escape. I'll always remember the Orange Light District and Harvard Mansion. Thanks to the sages, Forrest and Scott of brews, Croath and Wats of music, Brian of MTB, and Micah of quasi-reality.

Jaclyn, you are everything.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Through science, humans have harnessed empirical knowledge to predict natural phenomena that influence life on earth. Tracking the lunar cycle was one of the earliest attempts at using long-term average weather statistics, or climate, to time the annual planting of crops. Wildfires and volcanoes, which remain difficult to predict today, thwarted those early almanacs through natural air pollution and climate forcing. Though humans had burned biomass fuel for centuries, population growth and urbanization during the Roman Empire, Middle Ages, and Renaissance reduced forest resources near cities. The higher energy density and availability of coal in Britain popularized its use for domestic heating. In 1306, Parliament petitioned King Edward I to outlaw the burning of seacoal in response to the buildup of atmospheric pollutants (Commission and Argyll, 1871). The industrial revolution increased dependence on coal around the word, thus enabling humans to exert major feedbacks on both air quality and climate. In response to over 4,000 deaths during London's Great Smog of 1952, parliament passed the 1956 Clean Air Act, the first modern and substantial air pollution control legislation. Around the same time, numerical weather prediction (NWP) with computers became feasible (Charney, 1955). NWP was one of the first important steps leading to quantitative predictions of anthropogenic influences on air quality and climate.

Over the following decades, the United States enacted its own Clean Air Act (1963) with amendments (1970, 1977, 1990) and the Air Quality Act (1967). Among other provisions, the laws authorized: research into techniques for monitoring and controlling air pollution, as well as studies of air pollutant emission inventories; the establishment of National Ambient Air Quality Standards

(NAAQS); increased authority to regulate air pollutants; programs for acid deposition control; and regulation of chemicals that deplete the ozone layer. The U.S. Environmental Protection Agency (EPA) was established in 1970. One of the EPA's primary responsibilities is to ensure state, tribal, and local agencies monitor their own air quality and adhere to the NAAQS. Atmospheric chemistry models can be used to forecast regional haze and ozone exceedance events as well as to attribute past events to specific types of human or natural activity. In addition to reliable NWP, air quality forecasts and hindcasts require descriptions of emissions, chemical transformation, and removal (Carmichael et al., 2008).

Weather forecasting is an initial value problem, where the specified initial conditions have a strong influence on the quality of the forecast. Early weather forecasts were initialized with whatever atmospheric observations were available at the time. Charney et al. (1950) manually interpolated sparse observations to a model grid. Although more sophisticated interpolation schemes became available (e.g., Gilchrist and Cressman, 1954; Barnes, 1964), modern global NWP models have 10^7 degrees of freedom to describe with only 10^4 to 10^5 non-uniformly distributed observations for a given 6 hour period (Kalnay, 2003). The observation deficiency is greater for finer resolution mesoscale models used to predict the passage of fronts and storms across continents. With certain regions of the globe lacking observations, it became apparent that one could propagate observations from locations of abundance to those of scarcity using the models themselves. Present-day operational forecasts perform a new one week or longer forecast every 6 hours. The initial conditions are found by simulating the previous 6 hours, called the "analysis cycle," using some form of data assimilation (DA) algorithm.

There are several approaches to DA, including Optimal Interpolation (OI) (Gandin, 1966), 3D variational (3D-Var) (Sasaki, 1970), Kalman filter (KF) (e.g., Ghil et al., 1981), ensemble Kalman filter (EnKF) (Evensen, 1994), ensemble Kalman smoother (EnKS) (Evensen and Van Leeuwen, 2000), and 4D variational (4D-Var) (Marchuk, 1974; Penenko and Obraztsov, 1976; Dimet and Talagrand, 1986). In general, 4D DA methods (including EnKS) account for time-varying model-observation mismatch throughout the analysis cycle. Typically, 4D-Var uses the model as a strong

constraint (Sasaki, 1970), but a weak-constraint approach that accounts for model error is also possible (Trémolet, 2006). Alternatively, the flawed perfect model assumption can be corrected by testing multiple initial guesses of control variables or using multiple model configurations. The strong constraint is provided by the forward model equations, while the tangent linear (TLM) and adjoint (ADM) models are used to calculate the influence of a set of control variables on the sum of squared residual errors between model and observation, weighted by the model-observation error covariance matrix. This process is described in Chapters 2 and 3 for the special case of incremental 4D-Var as proposed by Courtier et al. (1994). While the model forecast error starts out large and is reduced, a second measure of error is the residual between the initial set of control variables and the optimized ones, weighted by the inverse of the background error covariance matrix. The benefit of this approach is to devise a solution for the forecast initial conditions that is physically consistent with the model physics and that accounts for the full-rank covariance between the variables in the state vector. The disadvantages of 4D-Var are that it requires developing an ADM and that it relies on a time-consuming iterative optimization. The need for an adjoint model can not be avoided, but the iterative process might be improved through algorithmic creativity.

In contrast to 4D-Var, ensemble-DA (e.g., Ensemble Kalman Filter (EnKF) and many variations (e.g. Evensen, 1994; Houtekamer et al., 1996; Anderson, 2003)) carries out multiple forward model simulations in parallel. Each additional ensemble increases the number of model modes that can be constrained while only adding wall-time in the form of gathering and distribution of data. Due to the random nature of the ensembles, which are much less numerous than the dimension of the state vector, ensemble methods require localization strategies to eliminate spurious state variable correlations outside a specified lengthscale. Hybrid methods that incorporate a full-rank background covariance matrix with the ensemble technique have improved ensemble-based forecasts (e.g. Hamill and Snyder, 2000; Lorenc, 2003; Buehner, 2005), but they also require a deterministic variational DA system running in tandem (Auligné et al., 2016). Hybrid methods have also been developed to incorporate stochastic information into adjoint-based deterministic forecast systems (e.g. Clayton et al., 2013), and more recently to meld stochastic and deterministic DA systems

seamlessly (Auligné et al., 2016). The attention paid to avoiding adjoint-based DA while reaping the benefits of deterministic methods stems from the incongruence of 4D-Var and short forecast turnaround times.

National forecast centers around the world perform six-hourly operational forecasts for up to a week or more ahead. Within that short six hour window, a fraction of the computational time is dedicated to using some form of DA to derive the best global and regional initial conditions and regional boundary conditions. If 4D-Var is used, the TLM and ADM integrations account for more than $\sim 95\%$ of the wall-time in each of ~ 100 sequential iterations in an optimization algorithm. The model resolution is often reduced in the TLM and ADM integrations (Zhang et al., 2014b) to reduce the single iteration wall-time, but doing so also reduces the amount of information that can be gleaned at fine scales. As the number of observations available to operational centers increases (e.g., from satellites), so does the potential for DA to constrain more modes of model variability. Just as additional ensembles in EnKF characterize additional modes, so do additional iterations in 4D-Var. Recent years have seen available computing resources moving toward increasing processor counts without increasing processor speeds, which does not bode well for sequential algorithms. In Ch. 4, we review some recent efforts to solve this problem, including a saddle formulation of weak-constraint 4D-Var (Fisher et al., 2011) and novel preconditioning strategies (e.g., Desroziers and Berre, 2012; Gürol et al., 2014) to reduce the number of iterations required. None of these has provided the same type of parallelism to 4D-Var that ensemble-DA enjoys.

While NWP has benefitted from a boon in observations, and continued development of more advanced DA methods, because of the large influence of weather on day-to-day life, atmospheric chemistry observations and modeling have lagged behind. The chemical observing network is not comprehensive, because of the vast number of chemical species to observe that have appreciable impacts on air quality and climate, and the higher cost of measuring pollutants as compared to wind, temperature, pressure, and water vapor. Particulate and ozone surface monitoring networks in major U.S. cities are used to enforce NAAQS. They are supplemented by columnar ozone and aerosol products from polar-orbiting satellite instruments (e.g., the Ozone Monitoring Instrument

(OMI) (Levelt et al., 2006), the Moderate Resolution Imaging Spectroradiometer (MODIS) (King et al., 1992), and the Multi-Angle Imaging SpectroRadiomenter (MISR) (Diner et al., 1998)). In addition to their use in operational monitoring, these observations and additional annual field campaigns are used to identify weaknesses in model and emission inventory descriptions. The latter is crucial since annual primary and precursor aerosol emissions have uncertainties anywhere between 7% and a factor of four, with larger variation on seasonal to diurnal scales for particular sectors (Streets et al., 2003; Suutari et al., 2001). Although in the case of operational forecasting the control variables are the initial conditions to the analysis cycle, in practice they can be any user-specified variable that influences the model forecast, such as boundary conditions, or surface fluxes of chemical species in the case of an atmospheric chemistry model.

As chemical observations become more readily available, Carmichael et al. (2008) suggested an increasing role for DA in solving the major problems facing air quality forecasting, including improving understanding of interactions between aerosols, dynamics, and clouds. Aerosols are liquid or solid particles suspended in a gas, which is the Earth's atmosphere for this application. There have been numerous studies identifying aerosols as short-term climate forcers through the semi-direct (Hansen et al., 1997; Koch and Del Genio, 2010) and indirect (Twomey, 1977; Lohmann and Feichter, 2005) cloud effects. These processes, in addition to direct radiative forcing, garner fine particulate matter a large, but uncertain impact on climate (Myhre et al., 2013). Modeling aerosol cloud interactions requires an online NWP-Chem model that integrates dynamic and chemical equations simultaneously. Chapter 2 lists several NWP-Chem models that might be used to predict aerosol-cloud feedbacks, and Appendix F includes a schematic of the typical processes described by them. The alternative, a chemical transport model (CTM), interpolates three to six hour meteorological fields from a separate NWP model. Grell et al. (2004) showed that vertical mass transport of chemical tracers is highly sensitive to the choice of online versus offline modeling methodologies due to variations in boundary layer mixing strength.

Using the Weather Research and Forecasting model with chemistry (WRF-Chem, Skamarock et al. (2008); Grell et al. (2005)), Saide et al. (2015a) showed that including the radiative and

microphysical effects of wildfire smoke in an NWP model could improve the forecasting of severe storm events. However, monthly regional fire emissions of both gas and aerosol mass remain uncertain by a factor of 10 or more around the world (e.g., Southeast/East Asia: Fu et al. (2012); Subharan Africa: Zhang et al. (2014a); California: Ch. 3). As model resolution goes up and the model domain gets smaller, the uncertainty of grid-scale emissions goes up even more. Using highly uncertain inventories in regional forecasting of concentration requires constant corrections to chemical initial conditions and/or surface fluxes through DA. Operationally, this requires expansive observational coverage that is only available from satellites. DA can also be used in research to diagnose the underlying errors in inventories using data from satellites or from a specific campaign, as we do in Ch. 3. Chemical DA in an online model is valuable for improving predictions of the aerosol impacts on health, regional climate, and weather.

WRF-Chem includes simultaneous (i.e., "online") simulation of gas and aerosol-phase chemical and thermodynamic processes alongside the atmospheric dynamics. WRF-Chem has been used to model aerosol-cloud interactions (e.g., Yang et al. (2011); Saide et al. (2012a); Makar et al. (2015)). Significant work has been done to provide a 4D-Var framework in WRFDA (Barker et al., 2005; Huang et al., 2009; Zhang et al., 2014b), and ADM and TLM versions of the NWP model components already exist (Zhang et al., 2013). Although WRF-Chem has already been used in sequential DA studies (see Ch. 2) to improve chemical initial conditions, these approaches have two limitations: (1) there are not enough observations, (2) chemical concentrations decay back to the emissions-driven values following the characteristic loss rate of each species. WRF-Chem has also been used in top-down emission studies through Lagrangian Particle Dispersion Modeling (LPDM) (see Ch. 3), which is useful when in-situ observations from surface sites or aircraft are available during periods of consistent convective boundary layer mixing. Several offline CTM-based inverse models can be used for constraining primary and precursor aerosol emissions (see references in Ch. 2) from in-situ or daily satellite observations, but they lack the necessary online feedbacks and spatial resolution necessary to predict total aerosol radiative forcing. Saide et al. (2015b) used WRF-Chem in an adjoint-free 4D variational method to constrain temporally varying smoke

emissions from a single fire. That method has also been used in operational smoke forecasting (personal communication, Pablo Saide), but is limited computationally to only solving for a small number of spatially aggregated source regions. Adjoint-based 4D-Var does not have the limitations of sequential DA, LPDM, or adjoint-free variational methods. To the author's knowledge, adjoint-based 4D-Var still has not been used in a regional NWP-chemistry model with online coupling to constrain grid-scale aerosol precursor emissions.

In order to answer the call to action by Carmichael et al. (2008) and to address processor scalability issues in adjoint-based DA, this work enables randomized incremental chemical 4D-Var in an NWP model with online chemistry (NWP-Chem). First, we create the foundation for a 4D-Var system that can probe the feedbacks between aerosols, dynamics, and microphysics, as well as constrain highly uncertain primary and precursor chemical emissions. Second, we apply the recently proposed Randomized Incremental Optimal Technique (RIOT) (Bousserez and Henze, 2016) in that same framework. WRF-Chem is selected for the NWP-Chemistry model, because of its widespread use, diverse options available for forward modeling, and existing NWP incremental 4D-Var capability. In this work, we create (Ch. 2) WRFPLUS-Chem and WRFDA-Chem, and use these tools to evaluate black carbon aerosol (BC) emission inventories. BC is used as a starting point for three reasons: (1) it is relatively inert, (2) it has highly uncertain natural and anthropogenic sources, and (3) its potential influence on climate covers a wide range, from large and warming to small and cooling (Bond et al., 2013). Ch. 2 describes the chemical TLM and ADM in WRFPLUS-Chem and an initial emission sensitivity study. Ch. 3 describes the specific steps of adapting incremental 4D-Var from NWP to gridded emission control variables. That chapter also includes a first application of WRFDA-Chem to the determination of optimal emissions of BC aerosol throughout California during the Arctic Research of the Composition of the Troposphere from Aircraft and Satellites in collaboration with the California Air Resources Board (ARCTAS-CARB) field campaign. Finally in Ch. 4, we apply RIOT in WRFDA-Chem to that same problem and assess its ability to reduce the wall-time of incremental 4D-Var and generate the optimal posterior control vector and posterior variance. We compare the results with an existing and widely used

sequential optimization algorithm.

Chapter 2

Development and application of the WRFPLUS-Chem online chemistry adjoint and WRFDA-Chem assimilation system

2.1 Introduction

Fine particulate matter impacts human health (Schwartz et al., 2007; Krewski et al., 2009) and climate (Myhre et al., 2013). Atmospheric climate forcing from aerosols is potentially large, but also highly uncertain owing to a complex spatial-temporal distribution of concentration, mixing state, and particle size for multiple species, each emitted from varying precursor sources, both anthropogenic and natural (Textor et al., 2006; Schulz et al., 2006). Depending on the species and quality of records, a nation's annual aerosol precursor and primary emissions have uncertainties anywhere between 7% and a factor of four, with larger variation on seasonal to diurnal scales for particular sectors (Streets et al., 2003; Suutari et al., 2001). Over these shorter time scales, aerosols impact meteorology through the semi-direct (Hansen et al., 1997; Koch and Del Genio, 2010) and indirect (Twomey, 1977; Lohmann and Feichter, 2005) cloud effects, which are both dependent on aerosol vertical profiles (e.g., Samset et al., 2013) governed by mixing.

Atmospheric models are used to improve our understanding of aerosol sources, distributions, and processes. Online numerical weather prediction and chemistry (NWP-chemistry) models integrate dynamic and chemical equations simultaneously, whereas offline chemical transport models (CTMs) interpolate meteorological fields from 3 to 6 hour reanalyses. Grell et al. (2004) showed that vertical mass transport of chemical tracers is highly sensitive to the choice of online versus offline modeling methodologies due to variations in boundary layer mixing strength. Additionally,

NWP-chemistry models account for moisture and temperature perturbations to dynamics due to aerosol microphysics and radiative forcing, while CTMs can not account for these feedbacks.

There are numerous online models with aerosol-meteorology feedbacks (e.g., WRF-Chem (Skamarock et al., 2008; Grell et al., 2005), COSMO-ART (Vogel et al., 2009), GEM-AQ (Kaminski et al., 2008),and IFS-MOZART (Kinnison et al., 2007; Flemming et al., 2009; Morcrette et al., 2009)). Better descriptions of sources, loss mechanisms, and vertical transport in coupled models are needed to increase accuracies in short-term climate modeling (Baklanov et al., 2014). To address this, chemical data assimilation can be used to improve short-term forecasts. In WRF, three-dimensional variational data assimilation (3D-Var) (Pagowski et al., 2010; Liu et al., 2011; Schwartz et al., 2012; Saide et al., 2012b, 2013), ensemble Kalman filter (EnKF) (Pagowski and Grell, 2012), and hybrid approaches (Schwartz et al., 2014) have all been used to improve chemical initial conditions. The limitation of these studies, using sequential methods, has been the decay of chemical concentrations back to the emissions-driven values following the characteristic loss rate of each species, necessitating periodic reinitialization with new observations. Using data assimilation solely to perturb initial conditions leaves behind underlying deficiencies in model description, emissions, or other input parameters.

In contrast to 3D approaches, 4D data assimilation attempts to minimize the discrepancy between model predicted values and observations at the same time observations are acquired. Variational 4D data assimilation (4D-Var) requires an adjoint, which calculates the sensitivity of a model metric to all input parameters, such as resolved aerosol precursor emissions. Several offline CTMs already have adjoints for constraining aerosol and aerosol precursor emissions, including GEOS-Chem (Henze et al., 2007), STEM (Sandu et al., 2005; Hakami et al., 2005), CMAQ (Turner et al., 2015), GOCART (Dubovik et al., 2008), and LMDz (Huneeus et al., 2009). Inverse modeling has been used to constrain aerosol emissions with 4D-Var, but only in offline models (e.g., Hakami et al., 2005; Dubovik et al., 2008; Henze et al., 2009; Wang et al., 2012). In addition to inverse modeling, derivatives calculated from CTM adjoints have been used to analyze sensitivities of model estimates to emissions (e.g., Turner et al., 2012). Online chemical 4D variational data assimilation

(4D-Var) has been performed with the global IFS-MOZART model, although without two-way coupling, to improve aerosol (Benedetti et al., 2009) and gas-phase (Inness et al., 2013) initial conditions. To our knowledge, 4D-Var still has not been used in a regional NWP-chemistry model with online coupling to constrain aerosol precursor emissions or other important model parameters, such as vertical mixing coefficients.

Here we present the first such system, building on existing capabilities of the WRF data assimilation (WRFDA) framework. WRFDA includes both 3D-Var (Barker et al., 2004) and incremental 4D-Var (Barker et al., 2005; Huang et al., 2009) algorithms, which are designed for constraining meteorological initial conditions (e.g., wind fields, temperature, moisture). For WRFDA v3.2 and later, WRF-4DVar requires calling the WRFPLUS forward (FWM), tangent linear (TLM) and adjoint (ADM) models. These models include adiabatic WRF dynamics, along with simplified surface friction (i.e., boundary layer), cumulus, and microphysics packages (Zhang et al., 2013). Here we integrate aerosol chemistry and vertical mixing from WRF-Chem into WRFPLUS, including complementary TLM and ADM components. While existing CTMs are capable of aerosol emission inversions, this development promises to introduce new insights into meteorology-chemistry couplings. We apply this system to black carbon (BC) aerosol, because of its important implications for climate (Bond et al., 2013) and health (Grahame et al., 2014). Additionally, the widespread use and development of WRF furthers the potential for continued model improvement and a community of future users.

BC is emitted from incomplete combustion of fuels. Major anthropogenic sources include residential cookstoves in developing countries, open crop burning, diesel transportation, and coal power plants with poor emission controls. Wildfires, or biomass burning, are the largest natural source. The major limitations to devising accurate bottom-up emissions inventories are poor activity data in developing countries and difficulty parameterizing complex biomass burning sources. Even in developed countries, changing economic landscapes affect real year-to-year emissions. Black carbon (BC) is unique among atmospheric aerosols as being radiatively absorptive, relatively inert, primary emitted, and having potentially complex cloud interactions. BC is possibly the second

most important human emitted pollutant in terms of climate forcing in the present-day atmosphere, with a net forcing of $+1.1 W m^{-2}$, but with 90% uncertainty ($+0.17$ to $+2.1 W m^{-2}$) (Bond et al., 2013). Also, reductions in BC emissions have been shown to reduce fine particulate health impacts (e.g., Anenberg et al., 2011).

The new TLM and ADM – referred to collectively herein as "AD/TL models" – aerosol treatments lay the groundwork for constraining aerosol precursor emissions using 4D-Var in a NWP-chemistry model. In Sect. 2.2, we describe the WRFPLUS-Chem and WRFDA-Chem model architectures. In Sect. 2.3, we describe the construction and verification of the AD/TL models of specific WRF-Chem forward model components. In Sect. 2.4, we describe a special checkpointing scheme that enables adjoint and tangent linear simulations longer than 6 hours, which are required for accumulating sensitivities of sparse chemical observations with respect to emissions. In Sect. 2.5, we demonstrate the capability of the adjoint model to calculate sensitivities of BC observation errors in WRFDA-Chem. Finally, we discuss future developments for WRFPLUS-Chem and WRFDA-Chem.

2.2 Methods

Creating the foundation for WRFDA-Chem required managing relationships between five related, but separate models. These include the (1) Weather Research and Forecast Model (WRF), (2) its "-Chem" variant, and the (3,4) WRFPLUS AD/TL models. Finally, (5) WRFDA 4D-Var requires communication of critical namelist and state variables to the FWM, TLM, and ADM. Figure 2.1 shows the relationships between these different models, including all AD/TL code that was previously developed, and code that we have added, modified, or plan to add.

2.2.1 Forward model

For this work, we use WRF version 3.6. The WRFPLUS-Chem code repository (`https://svn-wrf-model.cgd.ucar.edu/branches/WRFPLUSV3-Chem`) contains the most current version. Interested users can contact NCAR to request user access to the code. WRF contains multi-

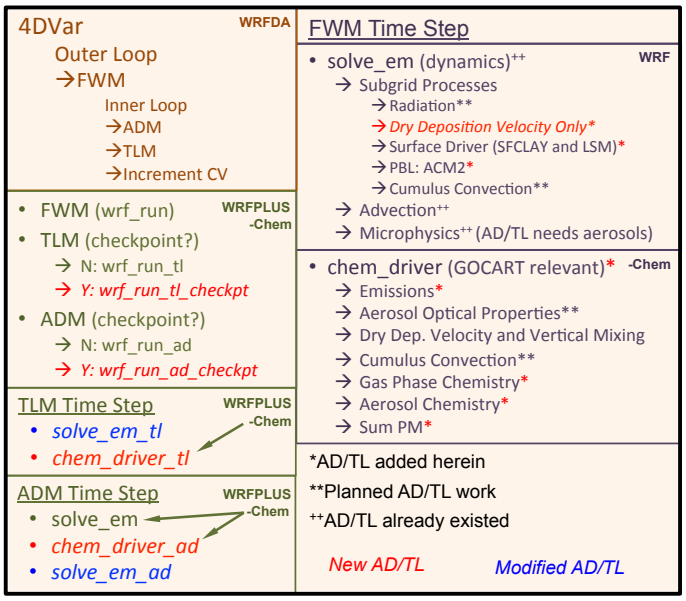

Figure 2.1: Dependencies between WRF, WRF-Chem, WRFPLUS AD/TL, and WRFDA. AD/TL development status is also noted.

ple non-hydrostatic dynamic cores and parameterization options for modeling unresolved physical processes. The FWM is identical in WRF and WRFPLUS, though typically only very simple unresolved physics are applied in WRFPLUS. In addition, WRF-Chem simulates the emission, deposition, transport, turbulent and cumulus mixing, wet scavenging, cloud interactions, and chemical transformation of trace gasses and aerosols. All of these processes are modeled at the same spatial and temporal resolution, which enables coupling WRF radiation and microphysics calculations directly with chemical processes.

The forward model configuration for which we have developed the corresponding TLM and ADM will be referred to as the "adjoint model configuration," because we use the same settings when running the adjoint. We use GOCART aerosols (chem_opt=300), wherein the chem array has 19 aerosol (e.g., SO_2, sulfate, black carbon, dust, sea salt) and zero gas-phase members. This option includes bulk mass sulfate chemistry and black carbon oxidative aging. We employ combined local and non-local ACM2 PBL mixing (Pleim, 2007a,b), with surface interactions handled by the Pleim-Xiu (PX) LSM (Xiu and Pleim, 2001; Pleim and Xiu, 2003; Pleim and Gilliam, 2009) and

surface layer (Pleim, 2006) mechanisms (all options seven). Soil moisture and temperature nudging are not used within the PXLSM. Prior to version 3.6, the WRF-Chem vertical mixing scheme solely carried out PBL mixing and dry deposition for chemical species. That vertical mixing depended on a (local) turbulent eddy mixing coefficient from a user-selected PBL scheme and a dry deposition velocity. There is new capability to calculate tracer turbulent mixing and dry deposition within the ACM2 subroutine itself, enabling non-local mixing. Trace gas and particle deposition velocities are calculated using characteristic resistances found using methods from Wesely (1989). Microphysics and radiation AD/TL models with aerosol feedbacks have not been incorporated into WRFPLUS-Chem yet. These crucial components will be partially adapted from previous work (e.g. Saide et al., 2012b, 2013), while others still need to be developed. Both microphysics and radiation are turned off for Sect. 2.3.3 verification simulations. In order to ensure appropriate radiative fluxes at the land-air boundary, the GSFCSW and Goddard LW radiation compute ground-incident radiation for the Sect. 2.5 adjoint sensitivity demonstration. However, online coupling between radiation and chemical species is deactivated.

2.2.2 Incremental 4D-Var

WRFDA uses an incremental 4D-Var method (Courtier et al., 1994) for finding the minimum of the cost function, J, by adjusting control variables (CV), $\boldsymbol{x}$. As described by Huang et al. (2009), the WRFDA cost function has three terms

$$J = J_b + J_o + J_c, \tag{2.1}$$

where J_b, J_o, and J_c are the background, observation, and balancing cost functions, respectively. J_c is not relevant to the current work. The background and observation cost functions are

$$J_b = \frac{1}{2}\left[\left(\boldsymbol{x}^n - \boldsymbol{x}^{n-1}\right) + \sum_{i=1}^{n-1}\left(\boldsymbol{x}^i - \boldsymbol{x}^{i-1}\right)\right]^\top \mathbf{B}^{-1}\left[\left(\boldsymbol{x}^n - \boldsymbol{x}^{n-1}\right) + \sum_{i=1}^{n-1}\left(\boldsymbol{x}^i - \boldsymbol{x}^{i-1}\right)\right] \tag{2.2a}$$

and

$$\begin{aligned} J_o &= \frac{1}{2}\sum_{k=1}^{K}\left\{H_k\left[M_k\left(\boldsymbol{x}^n\right)\right] - \boldsymbol{y}_k\right\}^\top \mathbf{R}_k^{-1}\left\{H_k\left[M_k\left(\boldsymbol{x}^n\right)\right] - \boldsymbol{y}_k\right\} \\ &\approx \frac{1}{2}\sum_{k=1}^{K}\left[\mathbf{H}_k\mathbf{M}_k\left(\boldsymbol{x}^n - \boldsymbol{x}^{n-1}\right) - \boldsymbol{d}_k\right]^\top \mathbf{R}_k^{-1}\left[\mathbf{H}_k\mathbf{M}_k\left(\boldsymbol{x}^n - \boldsymbol{x}^{n-1}\right) - \boldsymbol{d}_k\right]. \end{aligned} \tag{2.2b}$$

The background cost function is a penalty term, which ensures the departure of the posterior, $\boldsymbol{x^n}$, from the prior, $\boldsymbol{x^0} = \boldsymbol{x^b}$, remains within the bounds justified by the background error covariance, $\mathbf{B}$. The observation cost function measures the distance between the 4D-Var model solution, $\boldsymbol{x^n}$, and the observations, $\boldsymbol{y}$. M and H are the nonlinear model and observation operators, while $\mathbf{M}$ and $\mathbf{H}$ are their linearized forms, or tangent linear operators, used to propagate analysis increments $\boldsymbol{\delta x} = \boldsymbol{x}^n - \boldsymbol{x}^{n-1}$ from the earliest emission time to the k^{th} observation. $\mathbf{R}$ is the observation error covariance matrix. The innovation,

$$\boldsymbol{d}_k = \boldsymbol{y}_k - H_k\left[M_k\left(\boldsymbol{x}^{n-1}\right)\right], \tag{2.3}$$

is the residual error between the real and modeled observations k at the end of 4D-Var iteration $n-1$. This notation slightly differs from Huang et al. (2009), who employed K observation windows, each containing multiple observations.

For each iteration of incremental 4D-Var, the model is linearized about a trajectory, which is a collection of stored values of all model state variables at all time steps within the assimilation window. This trajectory enables propagation of sensitivities forward and backward in time within the TLM and ADM. Each of these models are called in an inner loop to calculate the gradient of the observation cost function, $\nabla_{\boldsymbol{x}} J_o$. An optimization algorithm uses the gradients to calculate optimal analysis increments to the CVs, which minimize the observation cost function. If the CVs, $\boldsymbol{x^n}$, depart too much from the initial guess for the current outer loop iteration, $\boldsymbol{x^{n-1}}$, the model

must be relinearized about the new state, $\boldsymbol{x^n}$, using M. The purpose of the two-level optimization is that approximating M with $\mathbf{M}$ transforms the cost function from a nonlinear to a quadratic form, and guarantees a unique solution $\boldsymbol{x^*}$ to the minimization (Courtier et al., 1994). Refer to Huang et al. (2009) for more details on the WRFDA incremental method, including a full expression for $\nabla_{\boldsymbol{x}} J$ given by Eqn. 7 of that article. The main purpose of this work is to introduce the AD/TL model components of WRFPLUS-Chem.

2.3 Tangent linear and adjoint model construction and verification

We have developed and tested adjoint and tangent linear code to represent aerosol-relevant processes in WRFPLUS-Chem. This development required a four step process:

(1) Automatically differentiate specific WRF-Chem modules using TAPENADE (Hascoet and Pascual, 2013) version 3.6.

(2) Verify standalone TLM and ADM derivatives against finite difference approximations; debug as necessary.

(3) Incorporate code manually into WRFPLUS.

(4) Repeat step 2 for fully integrated WRFPLUS-Chem model.

TAPENADE takes discrete Fortran or C source code as input, then generates either TLM or ADM code using a user-generated list of independent and dependent parameters. In addition to creating the differential code, TAPENADE reduces adjoint computational cost by eliminating unnecessary lines of code. Similar to Xiao et al. (2008) and Zhang et al. (2013), integrating the automatically differentiated adjoint code into WRFPLUS required significant manual intervention and debugging. Methods for constructing discrete adjoints are well-documented (Giering and Kaminski, 1998; Hascoet and Pascual, 2013). For the remainder of this section, we discuss the particular mechanisms for which we have created AD/TL models, and then we provide verification results for WRFPLUS-Chem.

2.3.1 Transport mechanisms

PBL physics and dry deposition in a column are handled by ACM2. The simple surface friction previously developed for WRFPLUS does not perform vertical mixing of tracers, which is a minimum requirement of any PBL scheme used in WRFPLUS-Chem. The ACM2 PBL depends on ground-atmosphere interactions that necessitate additional surface layer and land surface model (LSM) AD/TL code. For example, the ACM2 PBL scheme depends on the friction velocity U^* calculated in a surface layer scheme, which itself depends on wind speed, and the state variables u and v. ACM2 also depends on surface heat (HFX) and moisture (QFX) fluxes, which can be calculated within the surface layer or LSM code, but also depend on U^*. The dependence of HFX and QFX on ground-incident shortwave radiation (GSW) is calculated in the LSM. GSW is calculated in the radiation scheme, and depends on the aerosol composition and atmospheric moisture phase and distribution. Because we have not developed radiation AD/TL code, this coupling is not represented in WRFPLUS-Chem yet. The dependencies themselves are illustrative of how ACM2, and indeed most any other PBL scheme available in WRF, is appropriate for representing chemistry-meteorology interactions critical to understanding short-term climate impacts from aerosols. ACM2 is compatible with the Monin-Obukhov and PX (options 91 and 7) surface layer options, as well as the SLAB and PX (options 1 and 7) LSM options. TLM and ADM code is developed for all of these choices, and have been tested in standalone verification tests. In the interest of brevity, complete model verification in Sect. 2.3.3 has been limited to the two PX options.

Advection of inert tracers was added to WRFPLUS by X. Zhang (2012, personal communication). The same treatment has been applied to the "chem" array, with additional checkpointing and parallel communications. We generated standalone TLM and ADM code for deep cumulus convection as handled by the Grell-Freitas cumulus scheme (Grell and Freitas, 2014). One of the major benefits of this cumulus scheme is the ability to use online calculated cloud condensation nuclei (CCN) to account for the effect of aerosols on liquid and vapor water mass fractions. These parameters directly impact convection, including tracer transport. The ability of the standalone

AD/TL codes to produce the relevant members of the Jacobian has been verified for a single set of column conditions using similar methods as described in subsection 2.3.3. However, the FWM, TLM, and ADM do not yet account for vertical transport of chemical tracers, and thus have not been integrated into WRFPLUS-Chem.

2.3.2 Aerosol-specific components

GOCART is a bulk aerosol scheme that treats reactive species (BC, OC, sulphate) using a total mass approach and divides non-reactive species (dust, sea salt) into multiple size bins (Chin et al., 2000). Oxidative aging for both BC and OC is handled by a first-order decay from hydrophobic to hydrophilic forms using a time constant of 2.5 days. Sulphate (SO_4^{2-}) is produced from SO_2 and dimethyl sulfide precursor gases in GOCART. Sulphate chemistry also requires offline-calculated values for nitrate and OH radical, which are taken from climatologies available from the PREP-CHEM-SRC preprocessor (Freitas et al., 2011). WRFPLUS-Chem includes both the carbon and sulfate chemistry AD/TL codes, but only the BC component is tested and applied here.

Emissions of aerosol precursors in WRF-Chem is a linear process corresponding to specific chemistry and emission inventory options. Emission magnitudes are calculated, then distributed spatially and temporally, in offline preprocessors. Typically, emissions are read in hourly following some diurnal pattern. In order to make the emissions code easily differentiable, scaling factors are added to the emissions such that

$$\mathbf{E}_{c,i_{sc}} = \boldsymbol{\alpha}_{c,i_{sc}} \tilde{\mathbf{E}}_c. \tag{2.4}$$

At any simulation time, $\tilde{\mathbf{E}}_c$ are the emissions most recently read in from file for chemical species c. $\boldsymbol{\alpha}_{c,i_{sc}}$ and $\mathbf{E}_{c,i_{sc}}$ are the emission scaling factors and effective emissions, respectively, during scaling period i_{sc}. For emission inversions, the CVs, $\boldsymbol{x}$, are spatial-temporal resolved emission scaling factors. At the beginning of 4D-Var or during an adjoint sensitivity study, the scaling factors are set to unity. The scaling factors are applied in the FWM if environment option WRFPLUS==1 is set during compilation.

Dry deposition velocities are calculated in WRF-Chem within the dry deposition driver. In order to ease adjoint code construction and reduce checkpointing requirements, the dry deposition velocity calculation is moved to immediately precede the PBL driver as depicted in Fig.2.1. The new source code is similar to the dry deposition driver, except that only code corresponding to the GOCART aerosol option remains. The dry deposition AD/TL code accounts for dependencies of the dry deposition velocity on physical parameters (e.g., temperature, water vapor, U^*). As mentioned previously, the chemical concentrations are sensitive to dry deposition velocity within the PBL scheme.

2.3.3 Verification and linearity test

WRFPLUS FWM, TLM, and ADM performance were previously verified by Zhang et al. (2013). Here we use an alternative verification approach similar to that used by Henze et al. (2007). We use the TLM, ADM, and a centered finite difference approximation from the FWM to evaluate derivatives

$$\chi_{p,q} = \frac{\partial J_{p,f}}{\partial x_{q,0}}, \tag{2.5}$$

of some cost function at location p and time step f with respect to some CV at location q and the initial time 0. The finite difference derivatives are calculated from

$$\chi_{p,q}^{NL} \approx \frac{J_{p,f}\left(x_{q,0} + \delta x\right) - J_p, f\left(x_{q,0} - \delta x\right)}{2\delta x}, \tag{2.6}$$

where each evaluation of J results from a FWM evaluation with some perturbed value of $x_{q,0}$. δx varies between 0.1% and 10% of the value of $x_{q,0}$. The adjoint and tangent linear derivatives are found by forcing the model gradient fields, λ^* and λ, at J_p and x_q, respectively. The tangent linear gradient and adjoint gradient variables are analogous to state variables in the FWM. We force gradients of 1.0, indicating a 100% perturbation of the variable, and the resulting derivatives are retrieved from the model output gradient fields, such that

$$\chi_{p,q}^{TL} = \lambda_{p,f} = \mathbf{M}\left(\lambda_{q,0}\right) \tag{2.7}$$

and

$$\chi_{p,q}^{AD} = \lambda_{q,0}^{*} = \mathbf{M}^{\top}\left(\lambda_{p,f}^{*}\right), \tag{2.8}$$

where $\mathbf{M}^{\top}$ is the adjoint operator.

In order to evaluate our additions to WRFPLUS-Chem, we test cost functions equal to hydrophobic (BC_1) and hydrophilic (BC_2) black carbon concentrations in 100 different grid cells. We evaluate derivatives with respect to five state variables at three initial locations for each of those 200 cost functions. The CVs include initial conditions for BC_1, U, T, and Q_v, and also BC emission scaling factors, α_{BC}. All sensitivities apply over a 3 hour duration for a domain covering the southwest United States. For a full domain and model setup description refer to Sect. 2.5.1.1. Figure 2.2 shows that the maximum relative error between the TLM and ADM is in the 8^{th} significant digit. Thus we only need to compare the nonlinear model to the TLM to verify both the TLM and ADM. Those results are given in Fig. 2.3. The slope and R^2 statistic for a linear fit of those comparisons are very nearly unity for all CVs tested. Each of the plots in Fig(s). 2.2 and 2.3 depicts 600 derivative evaluations. A range of finite difference perturbations δx is used for U, T, and Q_v control variables in order to find a value of χ^{NL} with the best compromise between truncation and roundoff error. We test derivatives with respect to meteorological variables in order to show the AD/TL models will be functional in a setting with coupled chemistry and physics. In such a system, the emissions will impact meteorology, which in turn impacts concentrations. These results illustrate the capability of the AD/TL models to represent the latter part of that relationship. All of the verification results apply to a three hour simulation period, but longer simulations are needed to calculate the average influence of emissions on the modeled state-space.

2.4 Second order checkpointing

As discussed in Sect. 2.2.2, the nonlinear model trajectory is an integral component for propagating gradients in the AD/TL models. As one might imagine, the trajectory contains a large amount of information. WRFPLUS stores the entire double precision trajectory in memory in

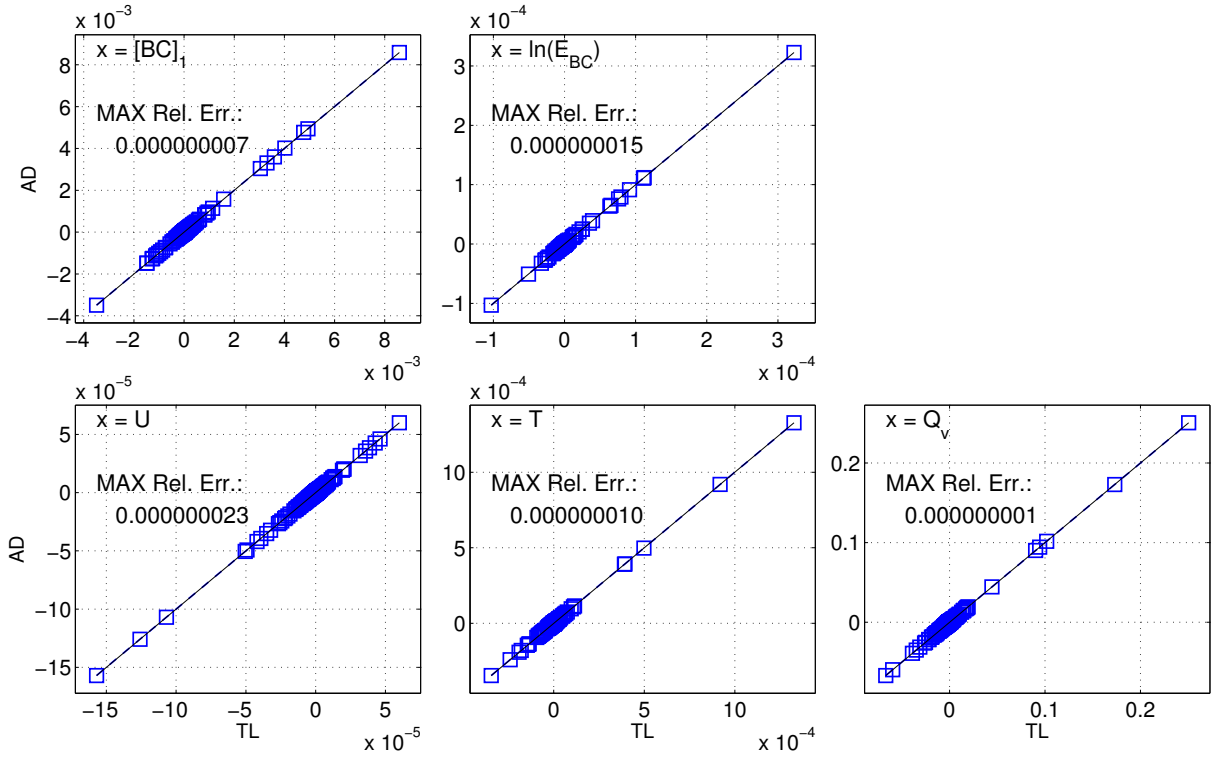

Figure 2.2: Comparison of ADM to TLM evaluations of $\frac{\partial[BC_1]}{\partial \boldsymbol{x}}$ and $\frac{\partial[BC_2]}{\partial \boldsymbol{x}}$ for 300 derivatives for each denominator variable.

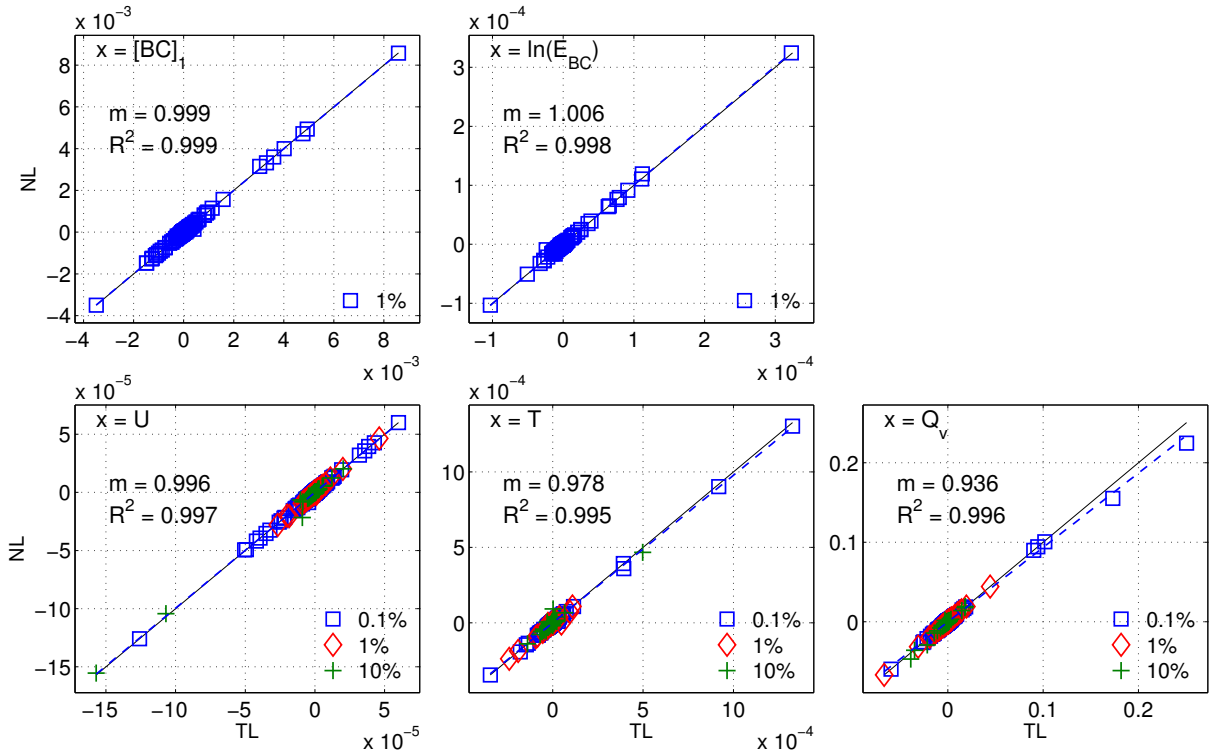

Figure 2.3: Comparison of nonlinear finite difference approximations to TLM evaluations of $\frac{\partial[BC_1]}{\partial \boldsymbol{x}}$ and $\frac{\partial[BC_2]}{\partial \boldsymbol{x}}$ for 300 derivatives for each denominator variable. The different markers for $\boldsymbol{x} = [U, T, Q_v]$ indicate the δx percentage that yielded a finite difference derivative closest to the tangent linear value.

order to eliminate expensive I/O time. This is very helpful with regards to storage, but presents a challenge in terms of memory. The system is designed for 6 hour operational assimilation windows. In a typical WRFPLUS-Chem simulation there are at least twenty-eight 3-dimensional state variables (8 physical, 1 to 3 moisture, and 19 GOCART species), and numerous other 2- and 1-dimensional state variables that must be included in the trajectory. For an illustrative domain, simulating 3 hours with a 90 second time step (18 km resolution), 79x79 columns, 42 levels, and a 5 cell boundary width, the trajectory would require 1.46GB per core on 64 cores. This final cost per core includes a 50% storage growth per doubling of the number of cores. Because the trajectory is stored for all time steps, required memory scales linearly with simulation duration and the number of simulated chemical species. For multi-day and multi-week inversions, as is typical in non-operational chemical data assimilation, the memory requirements become impractical for most cluster computing systems.

To solve this problem we implement a second order checkpointing scheme that shares the storage burden between the hard disks and memory. In a standard WRFPLUS adjoint simulation, the FWM is called first in order to calculate the trajectory. The FWM integrates the nonlinear equations from the initial to the final time, and stores the model trajectory at each time step. The ADM integrates the transpose of the linearized model equations backward in time, and at each time step reads the trajectory previously stored by the FWM. This process is depicted as "1st-order checkpoint" in Fig. 2.4. Since the storage limitation is driven by the duration of a simulation, we break the simulation into smaller segments, while maintaining continuity in the adjoint derivatives. The checkpointed adjoint simulation begins with a full FWM simulation beginning at the initial time, t_0, and ending at the final time, t_f. WRF restart files are written at time intervals equal to the checkpoint interval, Δt_c. Once the simulation is completed, the FWM is restarted at initial time equal to $t_f - \Delta t_c$. During that simulation, the trajectory is stored in memory. The trajectory is then recalled in an adjoint simulation that proceeds backward toward the current initial time. The checkpoint system alternately calls the FWM and ADM until returning to t_0. The major hurdle to integrating this second order checkpointing system into WRF-4DVar is that the trajectory is no

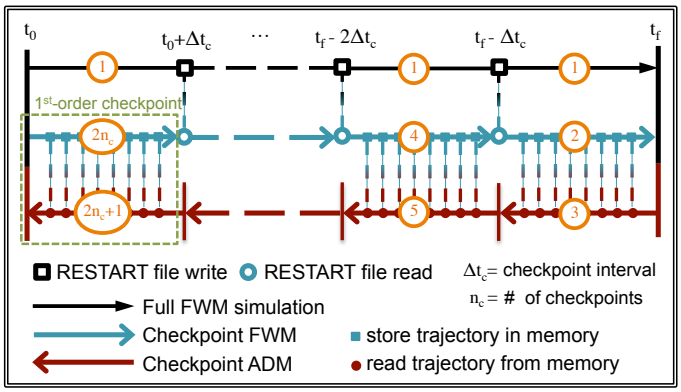

Figure 2.4: Second order checkpointing scheme implemented in WRFPLUS-Chem.

longer readily available to WRFDA for calculating modeled observations, $H_k\left[M_k\left(\boldsymbol{x}^n\right)\right]$, between the calls to the forward and adjoint models. Instead, these values must be calculated during either the full FWM (Step 1) or checkpoint FWM (Steps 2, 4, 6, etc.) simulations. We take the former approach. A similar checkpointing system is also implemented for the TLM in order to enable long duration incremental 4D-Var.

In order to ensure the checkpointing method delivers consistent derivatives to the non-checkpointed version, we again compare AD/TL derivatives to finite difference approximations. Because of the wall time required to calculate derivatives across extended time periods, we limit our tests to fourteen pairs of initial and final locations, q and p. For all of the J and x pairs tested in Sect. 2.3.3, the ADM and TLM agree to 13 or more digits over a 9 hour test. The improved performance relative to the previous 3 hour test came about after a few minor bug fixes. Because of this machine precision AD/TL agreement, we only compare the finite difference approximations to the TLM. For these checkpointed simulations, we analyze the derivative of a time variant cost function with respect to multiple control variables

$$\chi_{p,q}\left(t\right) = \frac{\partial J_p\left(t\right)}{\partial x_{q,0}}. \tag{2.9}$$

Doing so ensures that the derivatives are continuous across multiple checkpoint intervals and we are able to see the transient behavior of multiple finite difference perturbation sizes at times when

there are large discrepancies with the TLM. The finite difference approximations of derivatives of BC with respect to the physical variables grow more unstable with time. Thus, we calculate those derivatives only for a 6 hour period, while we test derivatives with respect to emissions for 48 hours. Here we also include derivatives of U, T and Q_v with respect to U and Q_v to ensure that those relationships are represented properly in the surface layer, LSM, and PBL AD/TL schemes, so that they may be used in a meteorological 4D-Var setting.

Figure 2.5 shows the resulting derivatives for nine different pairs of J and x for a single pair of q and p. Most importantly for multi-day 4D-Var emissions inversions, BC concentrations respond linearly to a 1% perturbation of emissions for at least 48 hours. Next, it becomes apparent why derivatives with respect to U and Q_v require multiple finite difference perturbation sizes to ensure one of them matches the TLM at a particular cost function evaluation time. There are times when either the smallest, largest, or no value for δx agrees with the TLM. However, the TLM has inflection points at the same times as the finite difference approximations, including during periods of intense oscillation, such as for $\frac{\partial U}{\partial U}$ and $\frac{\partial U}{\partial Q_v}$. The chemical concentrations respond nonlinearly to all U and Q_v perturbation sizes for periods longer than 1 hour in the plots shown, and longer than 3 hours for all test scenarios considered. Further testing of these coupled derivatives will be necessary to determine over what time period they are suitable for inverse modeling, and under what conditions the model nonlinearities cease to be a limiting factor. Future emission inversion work with coupled physics and chemistry will need to verify that $\frac{\partial J}{\partial \alpha}$ has a near linear response over the time frame considered. The behaviors noted here are consistent across the other thirteen pairs of q and p.

2.5 Sensitivities to BC emissions in California

Here we demonstrate the new WRFPLUS-Chem capabilities in an adjoint sensitivity study. For the present example, the 4D-Var cost function is the model response metric and the biomass burning, and weekday and weekend anthropogenic emissions are the model parameters of interest. This framework is used to analyze where and when these parameters most impact the model

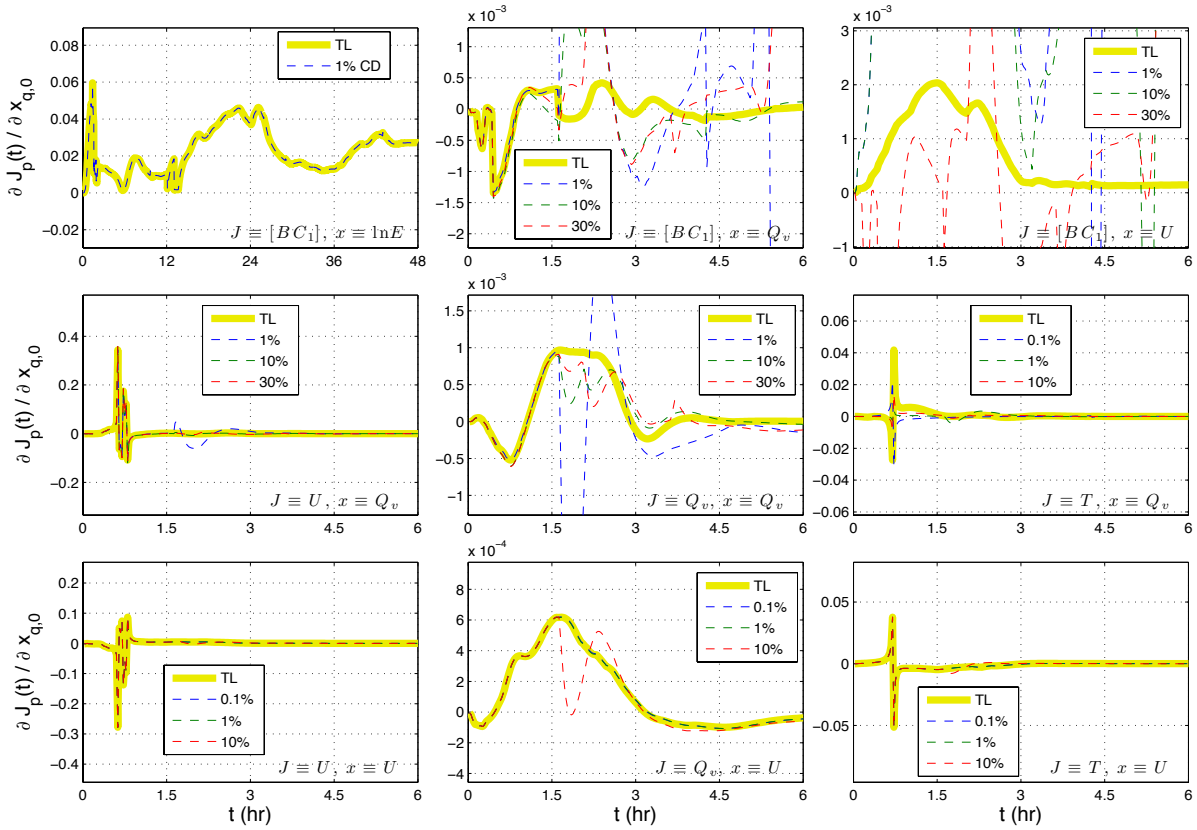

Figure 2.5: Time variant sensitivities of cost function J with respect to control variable x for multiple perturbations and the TLM with second order checkpointing.

performance and are thus in need of improvement.

2.5.1 Approach

For this demonstration, we calculate the sensitivity of the 4D-Var cost function in the first iteration. The background term is zero and there has been no prior CV increment (i.e., $\boldsymbol{\delta x} = \mathbf{0}$). Therefore, the cost function, Eq. 2.1, simplifies to

$$J = \frac{1}{2}\sum_{k=1}^{K}\{H_k\left[M_k\left(\boldsymbol{x}_b\right)\right] - \boldsymbol{y}_k\}^{\top}\mathbf{R}_k^{-1}\{H_k\left[M_k\left(\boldsymbol{x}_b\right)\right] - \boldsymbol{y}_k\}, \tag{2.10}$$

All off-diagonal covariances in $\mathbf{R}$ are assumed to be zero in order to enable timely matrix inversion.

2.5.1.1 Model configuration

The model domain encompasses California and other southwest U.S. states from 20 June 2008, 00:00:00 UTC to 27 June 2008, 09:00:00 UTC. We generated chemical initial conditions by running WRF-Chem for five days prior to the adjoint time period. We used the default WRF-Chem boundary condition for BC concentration of 0.02 $\mu\mathrm{g\,kg^{-1}}$. This is consistent with a single upwind Pacific ocean transect taken during the June 22 flight. Meteorological initial and boundary conditions are interpolated from 3 hour, 32 km North American Regional Reanalysis (NARR) fields. The horizontal resolution is 18 km throughout, and there are 42 vertical levels between the surface and model top at 100 hPa. The eta levels are 1.000, 0.997, 0.993, 0.987, 0.977, 0.967, 0.957, 0.946, 0.934, 0.921, 0.908, 0.894, 0.880, 0.860, 0.840, 0.820, 0.800, 0.780, 0.750, 0.720, 0.690, 0.660, 0.620, 0.570, 0.520, 0.470, 0.430, 0.390, 0.350, 0.310, 0.270, 0.230, 0.190, 0.150, 0.115, 0.090 , 0.07 , 0.052, 0.035, 0.020, 0.010, and 0.000. For a column where the ground is at sea level, there are 13 levels below 1 km and an additional 5 levels below 2 km. The subgrid physics options used are described in Sect. 2.2.1.

Anthropogenic emissions are taken from the U.S. EPA's 2005 National Emissions Inventory (NEI2005). Fire emissions are provided by the Fire INventory from NCAR (FINN Version 1) (Wiedinmyer et al., 2011, 2006). FINN uses Moderate Resolution Imaging Spectroradiometer

(MODIS) active fire locations and radiative power from NASA Terra and Aqua satellites, as well as speciated emission factors for four vegetation types, to calculate daily total 1 km resolution emissions. Burned areas are scaled to the combined fractional coverage of each 1 km^2 fire pixel by tree and herbaceous vegetation types assigned by the MODIS Vegetation Continuous Fields product (Hansen et al., 2003). Repeated fire detections in a single fire pixel are removed according to Al-Saadi et al. (2008). Plume rise injection heights are calculated in WRF-Chem by an embedded one-dimensional cloud-resolving model (Freitas et al., 2007, 2010; Grell et al., 2011).

2.5.1.2 Model-observation comparison

We compare the model to observations in individual time steps, which differs from previous data assimilation approaches with WRF. In the standard WRFDA 4D-Var architecture, observations are binned over intervals, or windows, typically of one hour or longer duration. Whereas WRFDA typically has k observation windows, here WRFDA-Chem and WRFPLUS-Chem handle k observations possibly each at a different time. In order to reduce memory requirements, the adjoint forcing is stored in a column array, instead of the 2D and 3D arrays that were required for each state variable for each window, k in WRFDA. Also, while WRFDA includes meteorological observation operators to be called offline, a fine temporal resolution observation operator must be called directly within WRFPLUS. The traditional approach made communication between WRFDA and WRFPLUS less cumbersome, but also limited the ability to use dynamic observations recorded across broad temporal scales in an inversion.

In-situ observations were collected throughout California during the June 2008 portion of the Arctic Research of the Composition of the Troposphere from Aircraft and Satellites field campaign in collaboration with the California Air Resources Board (ARCTAS-CARB) (Jacob et al., 2010). Instruments aboard the DC-8 aircraft measured trace gas and aerosol concentrations over four days, including elemental carbon (EC) from the single particle soot photometer (SP2) at 10s intervals (Sahu et al., 2012). Additionally, 41 Interagency Monitoring of Protected Visual Environments (IMPROVE) sites measured daily average surface light absorbing carbon (LAC) on June 20, 23,

and 26 by thermal/optical reflectance (TOR) analysis of quartz filters (Malm et al., 1994). Surface and aircraft observation locations during the campaign are indicated in Fig(s). 2.6 and 2.7. The aircraft trajectories are overlaid on MODIS Aqua true color images (Gumley, 2008), and locations of MODIS active fires (NASA).

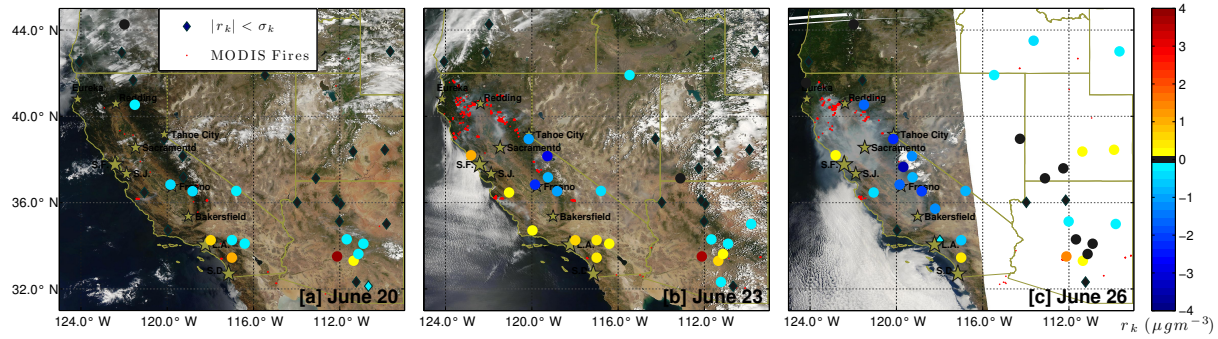

Figure 2.6: Surface site residual model error, r_k, overlaid on MODIS Aqua true color images and active fire retrievals. Observations with a bias less than one standard deviation are also indicated.

The observation operators for aircraft and surface observations require temporal averaging. The 10s resolution ARCTAS observations of BC concentration, pressure, latitude, and longitude are averaged to the 90 sec model time step, which is approximately the time the DC-8 would take to traverse a single 18x18 km^2 column. However, the 10s resolution ARCTAS BC concentrations are revision 2 (R2), while a later revision 3 (R3) product was released at 60s resolution only. The later revision includes additional mass in the 50-900nm size range as a result of applying a lognormal fit. In order to utilize this improved product, as well as leverage the finer resolution observations, the 10s BC mass is scaled by the mass ratio between the 60s R3 and the 60s average R2 datasets. The scaled 90s average observations are compared directly with the nearest model grid cell so that the model values are not interpolated. The pressure measurements are compared to online model pressures to determine the model level of each observation. For 24-hour average surface measurements from IMPROVE, the observation operator averages the nearest model surface grid cell concentration over all time steps within the observation period. For the few surface sites that have two air samplers simultaneously measuring, they are averaged together to prevent nonzero

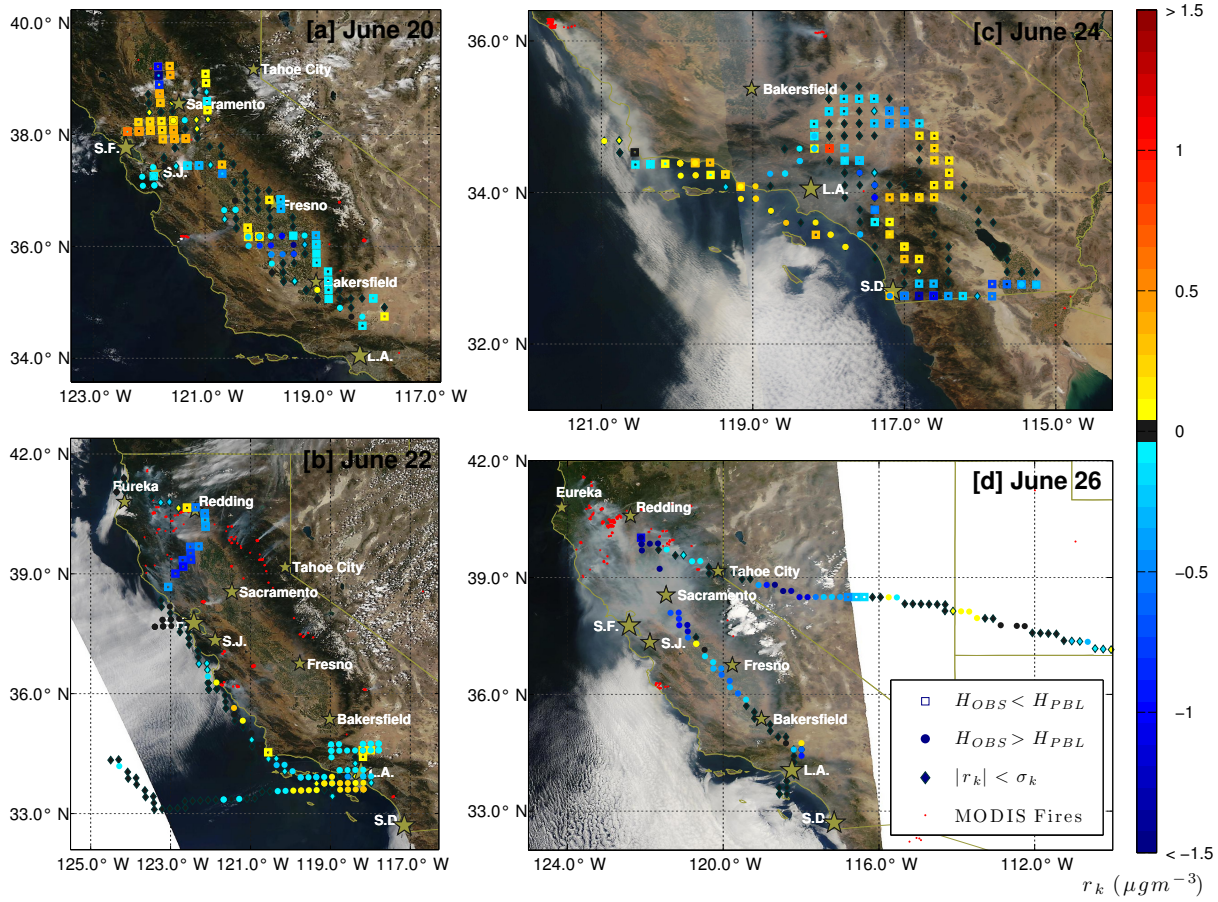

Figure 2.7: Aircraft residual model error, r_k, with indication for the observation height relative to the model PBL height overlaid on MODIS Aqua true color images and active fire retrievals. Observations with a bias less than one standard deviation are also indicated.

correlation in the cost function (i.e., off-diagonal terms in $\mathbf{R}$). After all averaging, there are 995 aircraft observations and 107 surface observations.

As depicted in Fig. 2.8, the WRF-Chem simulation is, on average, biased low for both the surface and aircraft observations. The lowest biased aircraft observations tend to be at higher altitudes, although this is not true in all cases. There are many high biased observations, and they tend to be at lower altitudes and to occur earlier in the simulation period when anthropogenic emissions dominate. Both surface and aircraft model predictions exhibit a wide spread of positive and negative errors. In order to determine potential causes for bias in specific locations, we consider the model residual errors, or simply "residuals,"

$$r_k = H_k \left[M_k \left(\boldsymbol{x}_b \right) \right] - \boldsymbol{y}_k, \tag{2.11}$$

for each aircraft observation k. Fig. 2.7 shows the statistically significant ($p < 0.32$) residuals for observations above and below the top of the model PBL. Section 2.5.1.3 describes relevant measures of observation variance and statistical significance.

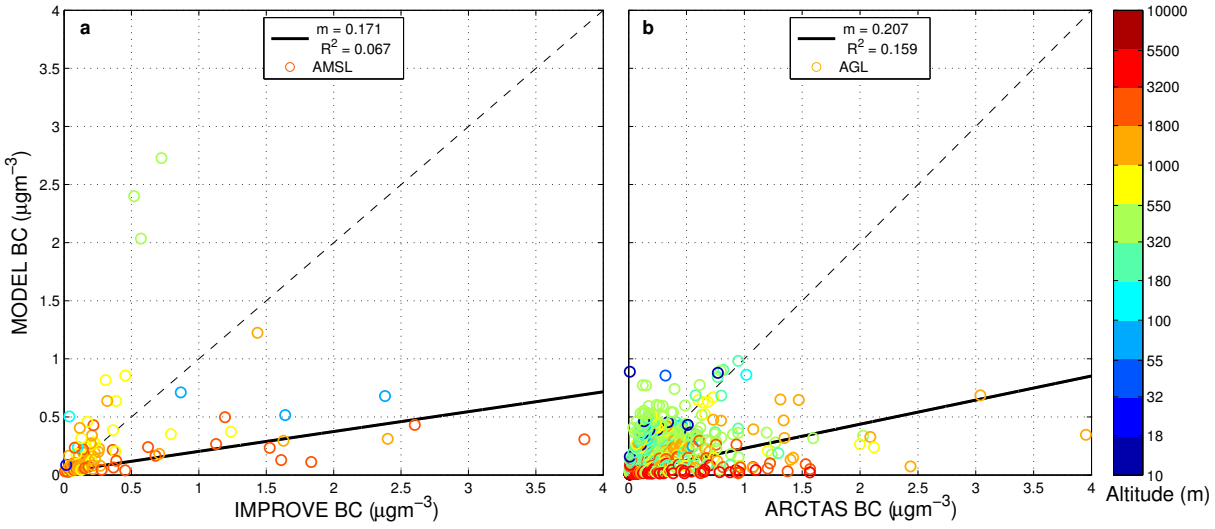

Figure 2.8: Linear fits between model BC concentrations with slope m and coefficient of determination R^2 for (a) IMPROVE surface and (b) ARCTAS-CARB aircraft observations colored by model height above mean sea level (AMSL) and above ground level (AGL).

Negative residuals, and hence low model bias, are most prevalent in northern California on

June 22 and 26, most likely due to under prediction of biomass burning sources. There is also low bias above the PBL in the southern San Joaquin Valley on June 20 and below the PBL inland from San Diego on June 24. Although neither case has visual smoke in the MODIS images, there were fires detected within 300 km. The largest positive residual occurs in Palmdale, CA close to landing on June 24. It could be indicative of either an emission error or the coarse horizontal resolution that collocates the airport with other significant nearby sources. Other notable high model biases aloft occur near cities during the flights on June 20, 22, and 24. Similarly, surface site biases are higher near cities, and along the coast. As might be expected, proximity to sources is a strong indicator of error magnitude, as that is where the highest concentrations occur. The error sign appears to be consistent above and below the PBL where such observations are collocated. Still, the spatial error pattern could reflect some combination of meteorology and emissions deficiencies. For the positive residuals off the coast of Los Angeles on June 22 and 24, there could be errors in predicting vertical mixing associated with the land-sea circulation or predominant near-surface wind direction. Discerning errors caused by emissions from those caused by meteorological mechanisms would require a separate in-depth study.

2.5.1.3 Variance and residual error significance

When **R** is assumed to be diagonal, each residual in the 4D-Var cost function is weighted inversely proportional to the observation error variance. The form of the cost function is based in Bayesian statistics, with an aim of converging on posterior control variables in a maximum-likelihood sense. However, using the variance alone to weight the residuals may result in very large cost function terms for relatively small residual errors. As our interest in this study is to determine how errors in emission estimates may be leading to model bias, we wish to ensure the largest residuals have the greatest weight, while also accounting for differences in statistical significance of particular errors. Thus we define the diagonal terms of **R** as

$$R_{k,k} = \frac{w_k}{\sigma_{k,k}^2}, \tag{2.12}$$

where w_k is an additional weighting term and $\sigma^2_{k,k}$ is the variance.

The variance is comprised of components due to both observation and model uncertainty as

$$\sigma^2_{k,k} = \sigma^2_k = \sigma^2_{k,m} + \sigma^2_{k,o}. \tag{2.13}$$

The model variance at each observation location is found from an ensemble of N_c=156 WRF-Chem configurations during the modeling period. Each ensemble member, c, uses a different combination of PBL, surface layer, LSM, and longwave and shortwave radiation options. Also, there are configurations both with and without microphysics and cumulus convection. From the ensemble, we use the population of residuals at each observation, k, to calculate the model variance

$$\sigma^2_{k,m} = \text{MAX}\left(\sum_{c=1}^{N_c} \frac{(r_{k,c})^2}{N_c - 1}, MML^2\right), \tag{2.14}$$

where *MML* is the minimum model limit. The minimum possible modeled BC concentration is limited by the boundary condition, which fills the entire model domain during the five day warm-up simulation. The *MML* is simply taken as the minimum model concentration for all observation locations and all model configurations, and is found to be 0.01 μg m^{-3} and 0.02 μg m^{-3} for aircraft and surface measurements, respectively, after rounding to the observation precision.

The IMPROVE instrument variance combines both relative and absolute uncertainties, the latter of which arises due to the minimum detection limit (*MDL*) (UC-Davis, 2002). For a single filter analysis, the variance (in μg^2 m^{-6}) is

$$\sigma^2_{l_k,inst.} = \left[\frac{\sqrt{34^2 + [(1000)\,(0.07)\,y_{l_k}]^2}}{1000}\right]^2. \tag{2.15}$$

The sub-observation index l_k is useful at sites with more than one air sampler. When a site has data from multiple instruments in a single day, we take their average and combine their instrument variances as

$$\sigma^2_{k,o} = \sum_{l_k}^{L_k} \frac{\sigma^2_{l_k,inst.}}{L_k^2}, \tag{2.16}$$

where L_k is the observation count. We assume the IMPROVE measurements fully represent the encompassing grid cell, since all sites are in remote locations and the samples are averaged over a 24 hour period.

In contrast, the aircraft variance must capture the representativeness uncertainty associated with comparing the average of an entire model grid cell with an average of multiple short duration segments of a sparse aircraft transect. According to commercial literature for the SP2 device, it has an *MDL* of 0.01 $\mu g\,m^{-3}$, which we assume applies over the 10 sec observation interval used during the ARCTAS campaign. The observations available through the NASA ARCTAS data archive have a BC mass concentration uncertainty of ±30%. Although Sahu et al. (2012) report ±10% BC mass uncertainty, that range is given by Kondo et al. (2011), who state their results are applicable in regions not impacted by refractory organic compounds, such as from biomass burning sources. Because there are significant burning sources in this domain, we adopt the more conservative range. We utilize the instrument uncertainties in a definition for total observation variance with components due to both averaging and representativeness, such that for each average aircraft measurement,

$$\bar{y}_k = \sum_{l_k=1}^{L_k} \frac{y_{l_k}}{L_k}, \tag{2.17}$$

the total variance is

$$\sigma_{k,o}^2 = MDL^2 + \sigma_{k,avg.}^2 + \sigma_{k,rep.}^2. \tag{2.18}$$

Adding the minimum variance associated with the *MDL* prevents the total variance from trending toward zero for any particular observation. This is important when using the variance in the cost function to ensure that near zero observations–which have low variances–with small residuals do not dominate the inversion. The averaging variance is the variance of the y_{l_k}'s that makeup $\bar{y}_k$, which is an attempt to capture the spread of true concentrations in a model grid cell. In the case that there is only a single observation, the averaging uncertainty is taken as double the instrument uncertainty. Thus,

$$\sigma_{k,avg.}^2 = \begin{cases} \sum_{l_k=1}^{L_k} \left[\frac{\left(y_{l_k} - \bar{y}_k\right)^2}{L_k - 1} \right] & \text{if } L_k > 1; \\ \left(2\sigma_{k,inst.}\right)^2 & \text{if } L_k = 1 \end{cases} \tag{2.19}$$

For any time step where $L_k < L_{max} = 9$, there is an additional variance penalty proportional to

the sum of the individual instrument variances,

$$\sigma^2_{k,rep.} = \sqrt{\frac{L_{max} - L_k}{L_{max}}} \sum_{l_k=1}^{L_k} \frac{\sigma^2_{l_k,inst.}}{L_k^2}, \tag{2.20}$$

where

$$\sigma_{l_k,inst.} = \mathrm{MAX}\left(MDL, 0.3 \cdot y_{l_k}\right). \tag{2.21}$$

In order to motivate the weight, w_k, applied to each residual model error, let us consider the primary inputs to the adjoint simulation, which are the adjoint forcings

$$\begin{aligned} \lambda^*_{k,m} &= \frac{\partial J}{\partial c_k} \\ &= \mathbf{H}_k^\top \sigma_k^{-2} \{H_k\left[M_k\left(\boldsymbol{x}_b\right)\right] - \boldsymbol{y}_k\} \\ &= \mathbf{H}_k^\top \lambda^*_{k,o}. \end{aligned} \tag{2.22}$$

c_k is any state variable on which H_k depends and which $M_k\left(\boldsymbol{x}_b\right)$ predicts. For our purposes, the state variables are modeled BC concentrations. The adjoint of the observation operator, $\mathbf{H}_k^\top$ transforms the forcing from observation space ($\lambda^*_{k,o}$) back to model space ($\lambda^*_{k,m}$). Thus, the forcing in observation space is

$$\lambda^*_{k,o} = \frac{r_k}{\sigma_k^2}. \tag{2.23}$$

Observations with significant model bias would require the largest perturbation in control variables to alleviate, and would seem to inform the inversion process the greatest. However, they must also have low total variance to contribute to an inversion. Figure 2.9 shows the surface and aircraft standard deviations plotted versus residual error. Also plotted in that figure are one and two standard deviation zones, as well as lines of constant $\lambda^*_{k,o}$ for all $w_k = 1$. Any residual falling outside the 2σ zone is one where the combined model and observation standard deviation is small enough to say with 95% confidence ($p < 0.05$) that the residual error deviates from zero (i.e., the model and observation disagree). These statistically significant model errors indicate that some kind of inversion is worthwhile. The relative contributions of observation and model variances is in general proportional to the relative magnitudes of observed and modeled concentration. Thus,

model (observation) variation contributes to a large fraction of uncertainty in positive (negative) residuals.

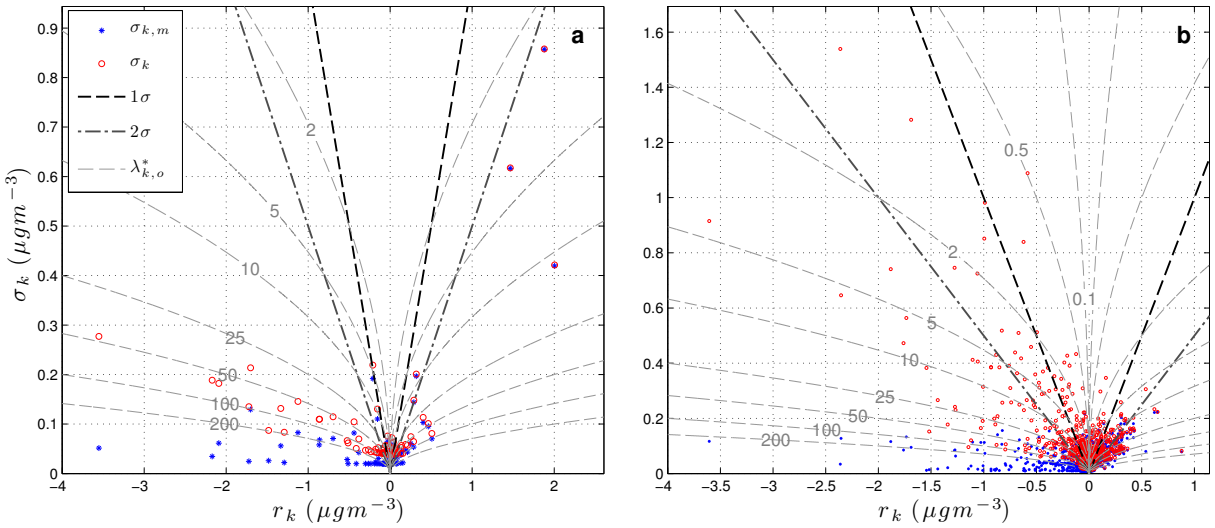

Figure 2.9: Model and total observation error standard deviation ($\sigma_{k,m}, \sigma_k$) versus model residual error (r_k) with adjoint forcing ($\lambda^*_{k,o}$) contours corresponding to $w_k = 1$ for (a) surface and (b) aircraft observations. 1σ and 2σ zones reflect regions of increasing statistical significance.

There are several outlier negative residuals with magnitudes much larger than the remainder of the population. A large portion of these have large enough uncertainty that their adjoint forcing is much less than that of other lower magnitude residuals. Consider the region where $|r_k|$ <0.5 μg m^{-3} and σ_k <0.3 μg m^{-3}. The adjoint forcing magnitude is between 10 and 200 μg^{-1}m^3, varying the mean forcing magnitude to the maximum for any observation in the whole population. The residual errors within the 1σ and 2σ zones are not statistically significant, yet they might have larger adjoint forcing than observations with larger residual error at higher significance levels. Applying these adjoint forcings as-is could drive the inversion to fitting data points with small absolute residual error. This adjoint forcing imbalance between high and low significance observations can be alleviated by a counteracting weighting scheme. In order to devise such a scheme, we consider which forms of statistical significance are important to this inverse problem.

Because our goal in an emission inversion is to reduce model bias by perturbing emissions, model bias is itself an important characteristic. We use the ensemble of model configurations to

calculate the variance in all residual errors, that is

$$\sigma_r^2 = \sum_{c=1}^{N_c} \sum_{k=1}^{K} \frac{r_{k,c}^2}{N_c K - 1}. \tag{2.24}$$

The residual standard deviations, σ_r, are 0.69 μg m^{-3} and 0.29 μg m^{-3} for surface and aircraft observation populations, respectively. After confirming that the residual errors are approximately normally distributed, the significance of the bias of a single observation relative to the entire population is

$$f_{POP,k} = \text{erf}\left(\frac{|\tilde{r}_k|}{\sqrt{2\sigma_r^2}}\right). \tag{2.25}$$

In statistics, the ratio of $\frac{|\tilde{r}_k|}{\sigma_r}$ is called the z-value, and denotes the number of standard deviations between $\tilde{r}_k$ and the expected value of zero. The variable $\tilde{r}_k$ indicates the user must select a specific form of residual error. Two examples are the mean or median of r_k. A third approach, and the one taken here is to use the residual found in the first 4D-Var iteration, $r_{k,n=0}$. $f_{POP,k}$ is a continuously variable p-value, or the percentage of the population of all $r_{k,c}$ that is less significant than $\tilde{r}_k$. Another measure of significance is visualized in the σ zones of Fig. 2.9, and was discussed previously. That is, for an individual residual error and variance, what is the probability that there will always be mismatch between the model and observation? The individual error significance is

$$f_{IND,k} = \text{erf}\left(\frac{|\tilde{r}_k|}{\sqrt{2\sigma_k^2}}\right). \tag{2.26}$$

The population and individual error significances are combined to derive the adjoint forcing weight,

$$w_k = \left[\left(f_{POP,k}\right)^{\gamma}\left(f_{IND,k}\right)^{1-\gamma}\right]^{\beta}. \tag{2.27}$$

The weighting scheme can be tuned for a specific application using the γ and β parameters to reshape the adjoint forcing contours. However, care must be taken when selecting γ, β, and $\tilde{r}_k$ to ensure convergence in 4D-Var to the mean of the Gaussian distribution of residual errors. Here we only introduce the weighting scheme and use it in a demonstration, but do not verify its validity. We use $\gamma = 0.5$ to provide some balance between the two measures of significance and $\beta = 2$ to ensure the weighting has a large impact. After calculating the w_k's according to Eqn. 2.27, the new

effective adjoint forcings are compared to the original values in Fig. 2.10. The weighting scheme is successful at reducing the impact of observation errors with low significance on the cost function.

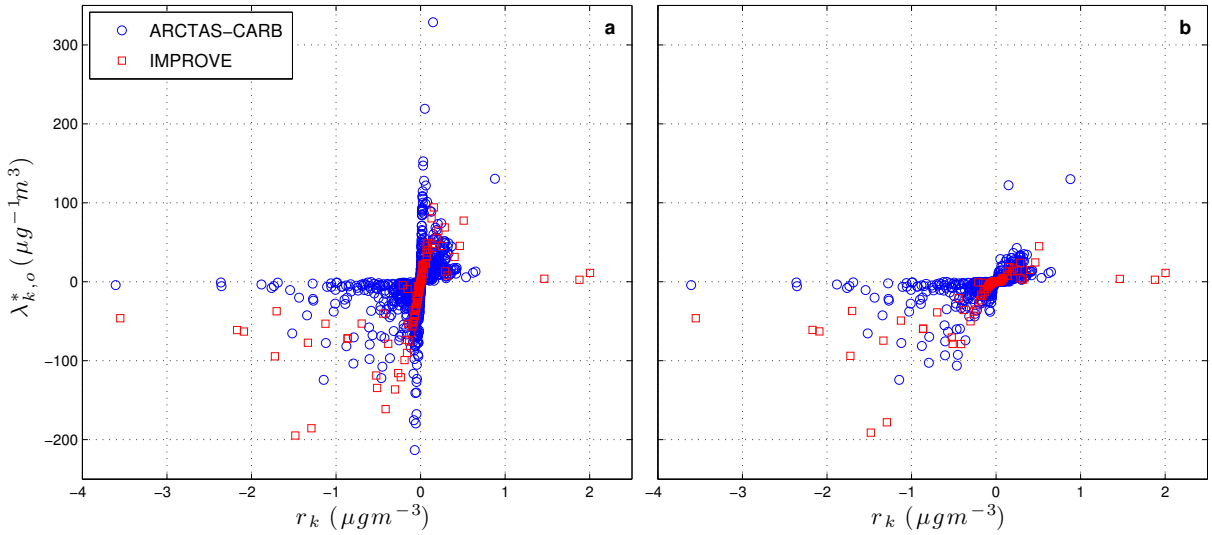

Figure 2.10: Adjoint forcing ($\lambda^*_{k,o}$) versus residual error (r_k) for ARCTAS and IMPROVE observations using weights of (a) w_k =1 and (b) w_k from Eqn. 2.27.

After applying the new weighting scheme, the $\lambda^*_{k,o}$ contours no longer converge on the y-axis as depicted in Fig.2.9. Instead, they exit radially from the origin in all directions. As both the population and individual z-values approach zero, the adjoint forcing converges toward

$$\lambda^*_{k,o} \approx \frac{r_k}{\sigma_k^2}\left(0.8\frac{|\tilde{r}_k|}{\sigma_r^{\gamma}\sigma_k^{1-\gamma}}\right)^{\beta} = 0.64\frac{r_k\tilde{r}_k^2}{\sigma_r\sigma_k^3}. \tag{2.28}$$

For our specific values of σ_r, all residual errors within the 2σ zone satisfy $|\lambda^*_{k,o}| \lesssim 5\ \mu\mathrm{g}^{-1}\mathrm{m}^3$ for surface, and $|\lambda^*_{k,o}| \lesssim 10\ \mu\mathrm{g}^{-1}\mathrm{m}^3$ for aircraft observations.

2.5.2 Results and discussion

With the weighting function applied, we calculate sensitivities of the 4D-Var cost function with respect to emissions for determining potential sources of model bias. The weights reduce the cost function from 5374 to 3784, which increases the normalized cost function sensitivity to emission

perturbations. Figure 2.11 shows fully normalized sensitivities,

$$\frac{\partial \ln J}{\partial \ln E_{i,j,d}} = \sum_{n=1}^{24} \frac{\partial \ln J}{\partial \ln E_{i,j,d,n}}, \tag{2.29}$$

for six days of the simulation. The sensitivity in a particular grid cell is summed over the local diurnal cycle for hours $n = [1, ..., 24]$ on day d. For anthropogenic emissions, the local time is calculated for discrete 15° time zones, whereas for biomass burning emissions, local time corresponds to the continuous sun cycle. Undoubtedly, there are locations with positive and negative sensitivities at different times of day that will cancel, but this temporally aggregated sensitivity is an attempt to obtain average daily relationships across the domain. Although the color bar has been saturated at $\pm 5{\cdot}10^{-3}$, the full range of sensitivities are from $-2.7{\cdot}10^{-3}$ to $+5.6{\cdot}10^{-3}$ and $-5.4{\cdot}10^{-3}$ to $+6.3{\cdot}10^{-3}$ for anthropogenic and biomass burning emissions, respectively.

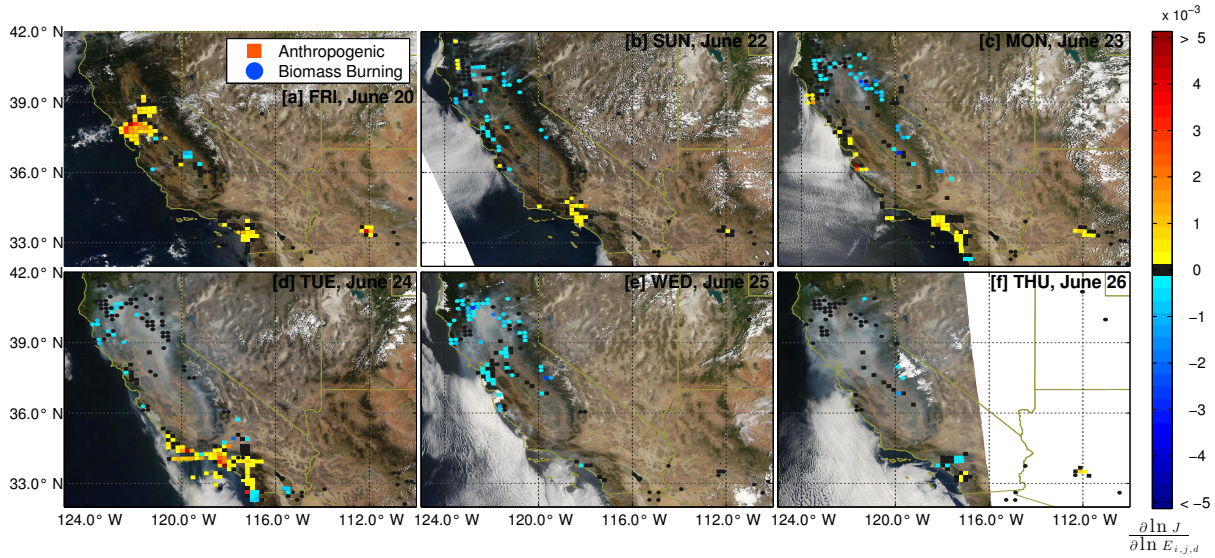

Figure 2.11: Normalized sensitivities ($\frac{\partial \ln J}{\partial \ln E_{i,j,d}}$) of the 4D-Var cost function (for surface and aircraft observations) with respect to anthropogenic and burning emission scaling factors overlaid on MODIS Aqua true color images for six days during the simulation. Anthropogenic sensitivities with magnitudes less than 1% of the maximum anthropogenic sensitivity magnitude are removed. There is a marker for all grid cells with non-zero burning emissions.

The magnitude of a normalized sensitivity corresponds to the fractional response in the cost function given a 100% perturbation of emissions in a grid cell. If the model were perfect, the

sensitivity magnitudes would be proportional to the difference between the background emission estimate and the true value. Thus a negative sensitivity indicates a location where estimated sources are too low, and vice versa. Because the sensitivities themselves depend on the emission magnitudes, they will change in each 4D-Var iteration, eventually converging on a minimum of the cost function where the sensitivities are zero. We use the sensitivities here as a qualitative indicator of emission errors, and not a quantitative conclusion as might be provided with a complete inversion.

The sensitivities exhibit a similar spatial-temporal pattern as the residual errors in Fig(s). 2.6 and 2.7, in general indicating that estimated anthropogenic emissions are too high, and that estimated fire emissions are too low. A more complex depiction of all BC emission errors arises in the sensitivities than the residuals alone might reveal. While most non-negligible burning sensitivities are negative, emissions from the Los Padres National Forest and northern redwoods on June 23 are potentially too high. The relative contributions of those fire estimates and simultaneous anthropogenic sources to positive coastal surface residuals near L.A. on June 23 are difficult to disentangle. However, the June 24 positive residuals from ARCTAS are more likely due to the anthropogenic sources. That is because the model transports smoke into the flight path of the DC-8 south of San Pedro, where BC concentrations are under predicted. Still, some anthropogenic source regions are under predicted as well.

The spatial variations in sensitivities are indicative of two phenomena. First, appreciable sensitivities will only arise in emissions that influence the particular observations available. Thus, full observation coverage is imperative to a successful inversion. Second, emission errors are heterogeneous in space and time. For biomass burning sources, heterogeneity arises due to missed detections in the MODIS active fire product, as well as potential errors in vegetation classification or attribution of a particular vegetation class to one of four land cover types used in FINN. Anthropogenic source error heterogeneity could be due to a static inventory from 2005 being used to describe emissions in 2008, or to spatial variations in BC emission factors for a particular source sector.

Comparative adjoint sensitivities are calculated using the SLAB LSM scheme (option 1) in

place of the PX option. In these results, the same positive coastal sensitivities are even more pronounced and widespread on June 23 and 24. Negative sensitivities to fires in the Sequoia and Inyo National Forests are larger in magnitude than those in the Sierras on June 23, but the spatial sensitivity patterns between SLAB and PX options are consistent on June 25. The differences are presumably due to changes in the residual error between the two configurations, since the weights and variances used are identical. $\tilde{r}_k$ was not recalculated for the SLAB case. The differing spatial sensitivity patterns indicate that the surface heat and moisture fluxes calculated by each LSM scheme contributes non-negligibly to the vertical mixing of BC to aircraft measurement altitudes.

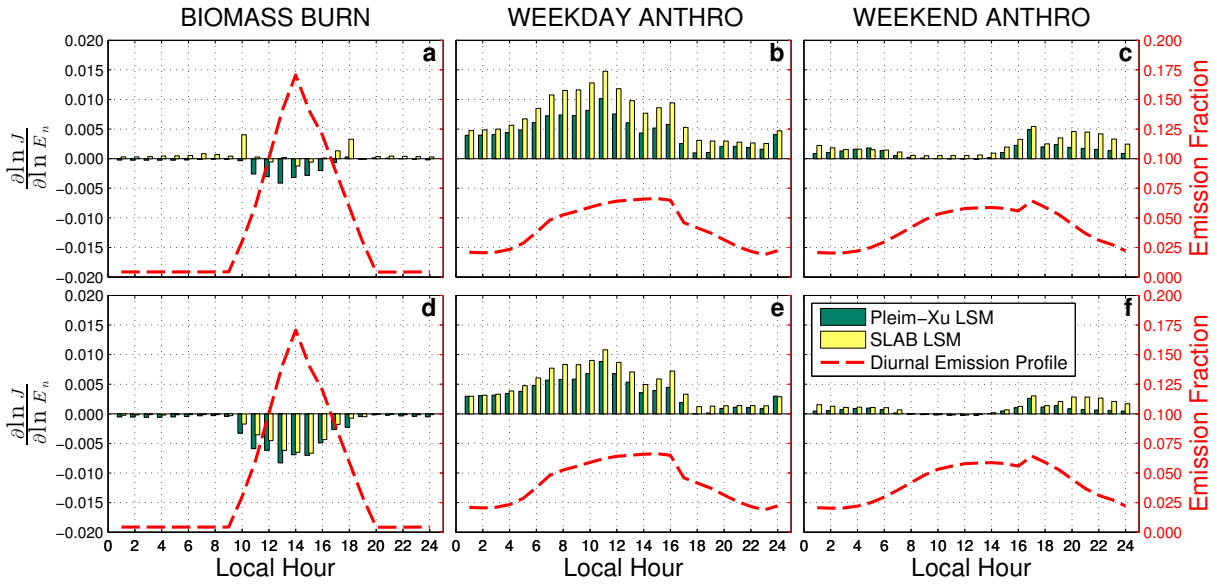

Figure 2.12: Diurnal normalized sensitivities ($\frac{\partial \ln J}{\partial \ln E_n}$) of the 4D-Var cost function with respect to emissions scaling factors for (a, b, and c) $w_k = 1$ and (d, e, and f) w_k from Eqn. 2.27. Also plotted are diurnal emission fractions. Sensitivities were calculated for two different WRF LSM options and are shown separately for biomass burning, and weekend and weekday anthropogenic emissions.

We also consider temporal sensitivity patterns to compare the two LSM schemes. Figure 2.12 shows the diurnal distribution of biomass burning, and weekday and weekend anthropogenic BC emission sensitivities for both of the LSM configurations, and for unity weights, $w_k = 1$ and w_k from Eqn. 2.27. Each bar in that plot represents a summation of sensitivities across the whole domain from 20 June 00:00:00 UTC to 26 June 23:00:00 UTC ($d = [1, ..., 7]$) within a particular

local hour, n, such that

$$\frac{\partial \ln J}{\partial \ln E_n} = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \sum_{d=1}^{7} \frac{\partial \ln J}{\partial \ln E_{i,j,d,n}}. \tag{2.30}$$

The signs and magnitudes of sensitivities fit the previous description for the spatially distributed temporal aggregation. The time period of emissions to which an observation is most sensitive depends on the altitude of that observation and the flow mechanisms that transport emitted aerosol mass to that observation. Thus, any conclusions drawn could be biased if observations do not have full temporal coverage, especially near sources. Since normalized sensitivities are proportional to emissions, it is to be expected that sensitivities at peak emission hours are magnified. Also, each hour of sensitivity is a sum of many diverse source locations. So while the net sensitivity in a single hour may be positive, the spatial distribution of sensitivities is much more varied, as was previously discussed.

The FINN biomass burning inventory applies an identical diurnal emission apportionment for all fires, regardless of vegetation, shading due to slopes, wind speed, or relative humidity. This scaling is applied in preprocessing. Both the PX and SLAB LSM setups seem to agree that the timing of the FINN burning emissions peak is correct within ± 1 hour, and that the peak should be sharper. Without the weighting scheme, the PX configuration indicates that burning emissions are too low in peak hours, while the SLAB configuration concludes that burning emissions are too high in off-peak hours. With the weighting scheme applied, both configurations agree that fire emissions need to be increased to reduce the cost function. The increased burning sensitivity magnitudes indicate the weighting scheme is successful at generating a cost function that is more robustly sensitive to emission perturbations. The relative disagreement in sensitivity magnitude between the two LSM configurations is attributable to differences in residual errors, r_k, and the resulting adjoint forcings, $\lambda^*_{k,o}$. Both configurations seem to agree that the timing of emissions is correct, and in fact the midday peak should be sharpened. However, the normalized sensitivities are proportional to emissions, meaning an emission peak should correspond to a sensitivity peak. Some work still needs to be done to interpret diurnal sensitivity patterns for use in a full inversion.

In contrast to FINN, NEI applies a variety of diurnal patterns to point, area, and traffic sources. The weekend and weekday emission profiles shown in Fig. 2.12 are the emission weighted averages for the entire domain. Individual sources may have a profile closer to flat, or alternatively zero overnight, and flat during daylight hours. The weighted average profile shown is close to the one used for commercial diesel traffic, since that is the largest BC source within the domain. Attributing sensitivities, or errors, to specific sectors is not straight-forward and doing so may require a smaller horizontal grid spacing to reduce the number of sectors per grid cell. Results for the weighted and unweighted cost functions are very similar. In general, anthropogenic emissions are too high throughout all times of the day on both weekdays and weekends. Both LSM configurations indicate the weekday profile peak should be sharper near 14:00LT, and not at 16:00LT, but also that emissions from 6:00LT to 16:00LT should be closer to the late evening and early morning magnitudes. The weekend sensitivities indicate the evening and morning emissions are too high, and that the daytime peak is timed about right, with the exception of the 18:00 LT spike. However, the relatively small magnitude of weekend sensitivities could also indicate there were not enough observations of anthropogenic sources on June 21 (SAT) and 22 (SUN) to draw definitive conclusions about emission timing.

Results for the two LSM options reveal the potential for model configuration to introduce bias in a 4D-Var inversion. For these particular observations, the posterior emissions from the PX option would likely be higher than those from the SLAB option, because of their relative sensitivity values. Model variability must be taken into consideration in 4D-Var sensitivity studies of high resolution emissions, because model variation represents a large fractional contribution to observation error variance for positive residuals, as shown in Fig. 2.9.

2.6 Conclusions

We have implemented, verified, and demonstrated the WRFPLUS-Chem coupled meteorology and chemical adjoint and tangent linear models for PBL mixing, emission, aging, dry deposition, and advection of BC aerosol. A second order checkpointing scheme enables tangent linear and

adjoint model runs longer than six hours. The adjoint was used in the first iteration of a 4D-Var inversion within WRFDA-Chem, where model-observation residual errors are compared for low- and high-temporal resolution IMPROVE surface and ARCTAS-CARB aircraft observations during one week of June 2008. A novel cost function weighting scheme was devised to increase the impact of high significance observations in future 4D-Var inversions. Results indicate that the weighting scheme is effective at generating robust sensitivities of the cost function to emissions. The adjoint sensitivities also indicate that anthropogenic emissions are over predicted and burning emissions are under predicted for the domain and time period considered. The diurnal sensitivities would seem to indicate that burning emission profiles should be steeper midday, while anthropogenic emission profiles should be flattened on weekdays and sharpened on weekends. A full inversion is necessary to quantify the magnitude of the errors in the emissions. Additionally, adjoint sensitivities found using two different LSM options indicate that the results of such inversions will be sensitive to the choice of model configuration.

The next steps are as follows. We intend to incorporate tangent linear and adjoint observation operators for useful remote sensing products (e.g., aerosol optical depth (Saide et al., 2013) and absorbing aerosol optical depth). This addition will enable WRFDA-Chem to be applied to a wider range of domains and time periods and operationally. The WRFDA-Chem optimization algorithm still needs to be applied to control variables for chemical species initial conditions and emission scaling factors. Future development and incorporation of radiation and microphysics adjoints (e.g. Saide et al., 2012b) will provide coupling between aerosols and meteorology, and provide new insights into sensitivities of direct, indirect, and semi-direct radiative forcing to emission sectors and locations. In addition to the aerosol applications discussed, WRFDA-Chem 4D-Var will also be suited to emission inversions for green house gases and other chemical tracers.

Chapter 3

Four dimensional variational inversion of black carbon emissions during ARACTAS-CARB with WRFDA-Chem

3.1 Introduction

Black carbon (BC) makes significant contributions to short term climate (Bond et al., 2013) and human health (Janssen et al., 2012) as a component of aerosolized fine particulate matter ($PM_{2.5}$) in the atmosphere. BC is emitted through incomplete combustion from natural and anthropogenic burning of biomass and fossil fuels. Open biomass burning (BB), which includes natural wild fires, deforestation, and agricultural waste and prescribed burning, accounts for 40% of total global BC emissions, while anthropogenic energy related sources (e.g., on- and off-road diesel and gasoline engines, industrial coal, residential cooking and heating) make up the remaining 60% (Bond et al., 2013). Future climate conditions that increase drought and fire prevalence (e.g., Spracklen et al., 2009) and increasingly regulated anthropogenic sources might lead to a reversal of these ratios in California (Mao et al., 2011) and globally (Jolly et al., 2015). In California, BB events have been shown to increase surface $PM_{2.5}$ concentrations by $\times 3$ to $\times 5$, compared to non-fire periods (Wu et al., 2006). The heterogeneity in BC emission and loss patterns and difficulty in replicating transport contribute to prediction uncertainty.

Despite the recognized importance of biomass emissions, large discrepancies remain in inventories in terms of biomass consumed and emitted chemical species. Zhang et al. (2014a) considered seven different inventories during February 2010 over Africa, which gave a range of $\times 12$ in total emitted OC and BC throughout the month. Fu et al. (2012) found similar variability between only

two inventories in Southeast and East Asia during January and April 2006. Zhang et al. (2014a) concluded that diffusion and loss mechanisms limit the effect of the monthly emission variability to only a ×2-3 range of monthly average domain-wide burden, AOD, and 2 m temperature. However, the inventory spread of emission magnitudes from larger sources led to column burden ranges of ×16-30 at the hourly to daily grid scales.

The large range in inventories at small scales results from the differing ways in which they are built. In order to be globally applicable, fire locating algorithms use remotely sensed hot spots from polar-orbiting satellites. Some provide additional regional locational and diurnal information with geostationary instruments. In all cases, daily emissions in a grid cell are calculated as the product of activity (kg burned) and emission factors for each species and vegetation class combination (kg emitted (kg burned)$^{-1}$). Bottom-up inventories combine rough estimates of burned area with vegetation densities and percent biomass burned associated with different Land Cover Types (LCT) to determine fire activity (e.g., Wiedinmyer et al., 2011; Reid et al., 2009; van der Werf et al., 2010). Top-down approaches use fire radiative power (FRP) measured by polar-orbiting or geostationary satellites and the LCT-specific energy content (e.g., Kaiser et al., 2012; Zhang et al., 2012), which circumvents using uncertain estimates of burned areas (Boschetti et al., 2004). A third approach combines the FRP with top-down constraints of aerosol optical depth (AOD) (e.g., Ichoku et al., 2012; Darmenov and da Silva, 2013). All three of these approaches cross reference fire locations with biome lookup tables to obtain the species-specific emission factors for each fire.

Improving short-term, local BC concentration predictions requires characterizing fine-scale spatial and diurnal patterns of BB emissions. The weakness of using only polar-orbiting data (e.g., Moderate Resolution Imaging Spectroradiometer (MODIS) instruments aboard Terra and Aqua) in bottom-up fire inventories is that there are nominally four overpasses per day, often with missed detections due to cloud- and smoke-cover or fire sizes beyond the instrument detection limits. Thus, these observations provide little information about the diurnal pattern of fire counts and FRP. Zhang et al. (2012) and Andela et al. (2015) devise methods for deriving climatological diurnal FRP patterns using geostationary observations. Both provide new information to modelers,

but the former is not generalizable to grid-scale diurnal variability and the latter precludes the possibility that diurnal FRP (Zhang et al., 2012) and emissions (Saide et al., 2015b) patterns may be bimodal for specific LCT's and fire regimes, or due to local meteorology.

In contrast to their BB counterparts, anthropogenic emissions of BC are periodic across weekly and annual time scales. Their spatial distributions are relatively well-known in developed countries, and less so in developing countries (Bond et al., 2013). Global estimates of annual anthropogenic BC emissions vary by $\times 2$ (Bond et al., 2013), national annual BC emissions in Asian countries and regions have uncertainties from $\times 2$ to $\times 5$ (Streets et al., 2003). In North America, including in California, uncertainties still persist in terms of characterizing the magnitude of emissions in a particular year, seasonal variability, and long term trends in activity and control strategies (Grieshop et al., 2006; McDonald et al., 2015). Bond et al. (2013) cite several inventories that give a range of $\times 1.7$ for annual U.S. anthropogenic BC emissions. However, like many other inventories, the U.S. EPA National Emission Inventory (Reff et al., 2009) does not specify uncertainty bounds either for the whole country or at state and county levels.

These challenges in characterization of both BB and anthropogenic emissions of BC and co-emitted species have led to the proliferation of top-down constraint methods of varying complexity and utility. Several studies have used adjoint-free methods for anthropogenic emissions in Los Angeles, California using aircraft measurements during the 2010 California Research at the Nexus of Air Quality and Climate Change (CalNex) campaign. Brioude et al. (2012) constrained CO, NO_x, and CO_2, and Cui et al. (2015) constrained CH_4; both applied a Lagrangian Particle Dispersion Model (LPDM). Peischl et al. (2013) constrained CH_4 using a mass balance approach and light alkane signatures from multiple sectors. LPDM benefits from being able to resolve sources on as fine of a grid resolution as is used in the underlying model. Both LPDM and mass balance are limited to linear tracer problems where observations are recorded under specific meteorological conditions. Wecht et al. (2014) used GEOS-Chem in an analytical inversion to compare constraints from the CalNex aircraft measurements with those from present and future satellite observations of CH_4 throughout California. Although an analytical inversion does not require an adjoint, the

approach is limited, computationally, to constraining only a few sources, which imposes aggregation error (Mao et al., 2015). Adjoint-based four-dimensional variational data assimilation (4D-Var) is able to account for nonlinear behavior between the emission sources and observation receptors by calculating exact gradients across physical processes. Such an approach does not have the limitations imposed by mass balance, LPDM, or analytical inversions, but does require development of an adjoint. The gradients are usually calculated through an adjoint model, although recent work (Saide et al., 2015b) performs 4D-Var on a limited area fire without an adjoint. That new approach, while easier to implement, is limited to solving for only a few spatially-distributed sources due to computational limitations.

In this study, we adapt the adjoint-based incremental four dimensional variational data assimilation (incremental 4D-Var) used in the WRFDA weather forecasting system (Barker et al., 2005; Huang et al., 2009) to solution of tracer surface flux estimation problems. We apply the resulting tool, WRFDA-Chem, to constrain anthropogenic and BB sources of BC throughout California during the Arctic Research of the Composition of the Troposphere from Aircraft and Satellites in collaboration with the California Air Resources Board (ARCTAS-CARB) field campaign. In June 2008, ARCTAS-CARB characterized aerosols and trace gases throughout California with DC-8 aircraft flights on 20 (Friday), 22 (Sunday), 24 (Tuesday), and 26 (Wednesday) June (Jacob et al., 2010). Sahu et al. (2012) used BC total mass measurements from a single-particle soot photometer (SP2) and other simultaneous gas-phase measurements to identify and characterize anthropogenic and BB plumes in California. We assess the capability of these same observations and every-3$^{\text{rd}}$-day surface measurements from the Interagency Monitoring of PROtected Visual Environment (IMPROVE) network to constrain errors in BC surface fluxes when used in 4D-Var. As described in (Guerrette and Henze, 2015), this approach of assimilating chemical tracer observations in a regional numerical weather prediction and chemistry model is unique in the context of previous 4D-Var flux constraints. We also estimate emissions, their associated uncertainties, and provide diagnostics for observing system evaluation at high spatio-temporal resolution (hourly, 18 km $\times$ 18 km). The approach taken in this work is described in Sec. 3.2, including the forward, adjoint,

and tangent linear models, the prior inventories and domain, and the incremental 4D-Var method implemented in WRFDA-Chem. Section 3.3 describes the application of WRFDA-Chem to the BB and anthropogenic emission inversion problem during ARCTAS-CARB. We conclude with a summary and recommendations for future measurement campaigns and emission inversion research.

3.2 Method

3.2.1 Nonlinear, adjoint, and tangent linear models

Incremental 4D-Var requires forward nonlinear (NLM), adjoint (ADM), and tangent linear (TLM) models. The NLM is nearly identical to WRF-Chem (Grell et al., 2005) with the addition of emission scaling factors. The GOCART option facilitates 19 species, including 4 gas and aerosol species for sulfate chemistry, hydrophobic and hydrophilic BC and organic carbon, 5 size bins for dust, 4 bins for sea salt, and 2 diagnostic species for $PM_{2.5}$ and PM_{10}. While we use GOCART, the results presented are limited to BC. The model configuration is the same as was used in Guerrette and Henze (2015), and is summarized as follows: ACM2 PBL mixing (Pleim, 2007a,b), Pleim-Xiu land surface model (Xiu and Pleim, 2001; Pleim and Xiu, 2003; Pleim and Gilliam, 2009) and surface layer (Pleim, 2006) mechanisms without soil moisture and temperature nudging, Wesely dry deposition velocities (Wesely, 1989), GSFC shortwave and Goddard long wave radiation, and microphysics turned off. Microphysical and radiative responses to online aerosols are also turned off, because they are not included in WRF-Chem for GOCART.

We utilize the recently developed WRFPLUS-Chem (Guerrette and Henze, 2015), which contains ADM and TLM code extending the original WRFPLUS software (Zhang et al., 2013). WRFPLUS-Chem describes chemical tracers in the context of planetary boundary layer (PBL) mixing, emissions, dry deposition, and GOCART aerosols. ADM and TLM gradients have been verified against finite difference approximations. Second-order checkpointing reduces the memory footprint to a feasible level for ADM and TLM simulations over longer durations ($>\sim$ 6 hr) and/or that use many chemical tracers ($>\sim$ 10). Guerrette and Henze (2015) applied the ADM in calculat-

ing sensitivities relevant to the emission inversion carried out here. Sec. 3.3.5 includes a comparison of the results of that study with the posterior emissions here.

The model domain is similar to that used by Guerrette and Henze (2015). The spatial extent encompasses California and other southwest U.S. states. We conduct two emission inversions, the first on 22 June with a focus on biomass burning sources, and the second on 23-24 June with a focus on anthropogenic sources. We generated chemical initial conditions by running WRF-Chem from 15 June 2008, 00:00:00 up until the beginning of each inversion period. We used the default WRF-Chem boundary condition for BC concentration of 0.02 $\mu g\,kg^{-1}$, which was found to be consistent with observations with an upwind flight on 22 June. Meteorological initial and boundary conditions are interpolated from 3 h, 32 km North American Regional Reanalysis (NARR) fields. The horizontal resolution is 18 km throughout 80×80 columns, and there are 42 vertical levels between the surface and model top at 100 hPa.

3.2.2 Prior emission inventories

The prior includes sources of BC from anthropogenic activity and natural wild fires. Anthropogenic emissions are taken from the U.S. EPA's 2005 National Emissions Inventory (NEI05) for mobile and point sources, including for example diesel on-road and power production from coal. The individual sectors are lumped together for each grid cell. We represent BB emissions using three different wild fire inventories, FINNv1.0 and v1.5 both at 1 km × 1 km resolution (Wiedinmyer et al., 2011, 2006), and QFEDv2.4r8 at 0.1° × 0.1° resolution (Darmenov and da Silva, 2013). FINNv1.5 is readily available through NCAR (`http://bai.acom.ucar.edu/Data/fire/`) to WRF-Chem users, while FINNv1.0 is no longer supported. However, we include FINNv1.0 in this study, because it shows equivalent value as a prior. FINN and QFED fall into the first (bottom-up) and third (top-down constraint with AOD) category of BB inventories described in Sec. 3.1, respectively. QFED scales global aerosol emissions from four biome types through multiple linear regression between observed MODIS aerosol optical depth (AOD) and modeled GEOS-5 AOD during the years 2004-2009. For temperate forests, which produce 80% of the wild-fire BC in California

during this modeling study (June 2008), QFED scales aerosols by $\times 4.5$ throughout the world. This global scaling is problematic for the California fires, because the GEOS-5 AOD is biased high in the Western U.S. during the summer fire seasons of 2006-2008 (Fig. C14 of Darmenov and da Silva, 2013). In order to match the regional climatological AOD scaling factors for the Western U.S., we scale all QFED BC sources by $\times \frac{1}{3}$. This scaling is already taken into account in the prior emissions shown in Sec. 3.3.3.

The WRF preprocessor distributed with the FINN inventory is used to distribute ASCII formatted lists of both FINN and QFED daily speciated fire emissions to hourly netcdf files readable by WRF. The diurnal profile follows the Western Regional Air Partnership profile WRAP (2005), and is defined by a flux peak from 13:00 to 14:00 Local Time (LT), and flat fluxes equal to 2.5% of the peak value between 19:00 LT and 09:00 LT. Through modeling experience, we found two bugs with how the FINN preprocessor interprets the WRAP profile and have fixed them for this case study. The total FINNv1.0 emissions across the model domain before and after fixing these bugs are plotted in Fig. 3.1 along with MODIS active fire counts (NASA). The first bug relates to how the timezone of a particular fire is calculated from longitude. The preprocessor converts a decimal longitude to integer time zone bins; this allows a fire at 120.1°W to be an hour earlier in the diurnal profile than a fire at 119.9°W, even though they should be at nearly identical positions in the WRAP profile. Such behavior might apply to anthropogenic emissions, where cities near time zone borders follow different daily cycles of activity, but not to natural activity related to the 15° per hour cycle of the sun.

The second bug, and the one most visible in Fig. 3.1, is in the redistribution of UTC fire detections in to LT emissions. MODIS Terra and Aqua overpass times are distributed around noon and midnight LT globally, with some adjustment as the image capture location moves farther from the equator. The fire hot spots are detected in UTC days, and their emissions are profiled according to LT periods corresponding to the same UTC day as the detection. In California, where the LT is UTC minus 8 hours, the noon overpass corresponds to 20:00 UTC, and 00:00 UTC corresponds to 16:00 LT on the previous day (sun cycle). Therefore, when a fire is detected during nearly

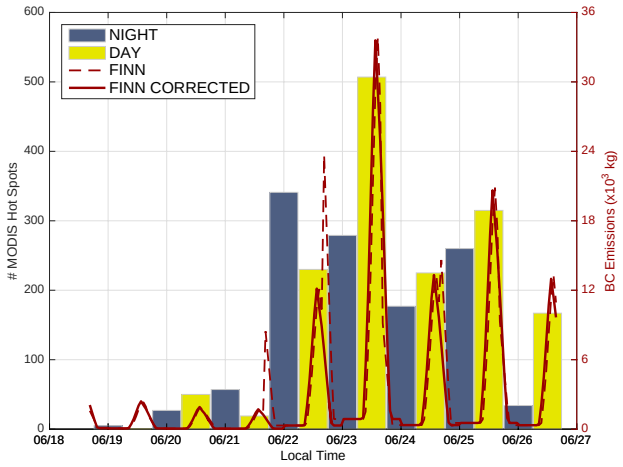

Figure 3.1: MODIS fire hot spot detections, excluding those with confidence less than or equal to 20% and double detections within 1.2 km of each other (left axis) and domain-wide FINNv1.0 BB emissions during the ARCTAS-CARB campaign, with and without fixes described in Sec. 3.2.2 (right axis).

peak heat and emission fluxes at noon, a large fraction of the flux is apportioned to the previous afternoon. For locations east of the International Date Line, the LT reallocation is in the opposite direction. In either case, some portion of the profile is shifted by 24 hours. This error is apparent as a temporal discontinuity in the case of transient fires that vary significantly in magnitude from one day to the next, especially after a recent ignition. Since the domain used here is nearly confined to a single time zone, we simply move the emissions forward one day for times between 16:00-23:00 LT (00:00-07:00 UTC). A more robust fix will need to be implemented in a future preprocessor.

Another error in the prior BB emissions is less easily resolved. Figure 3.2 shows where the MODIS active fires are located relative to the inventory fire locations. Since QFED fires are provided on a LAT-LON grid, the fire centers do not coincide with its grid centers. When the inventory is distributed to the 18 km model grid, some emissions are shifted over by one column relative to the FINN locations. There are several additional spurious emission locations in QFED, where no active fires were detected on either 21 or 22 June. In a month-long simulation, differences in fire gridding between several inventories can be averaged out. In the shorter term inversions over

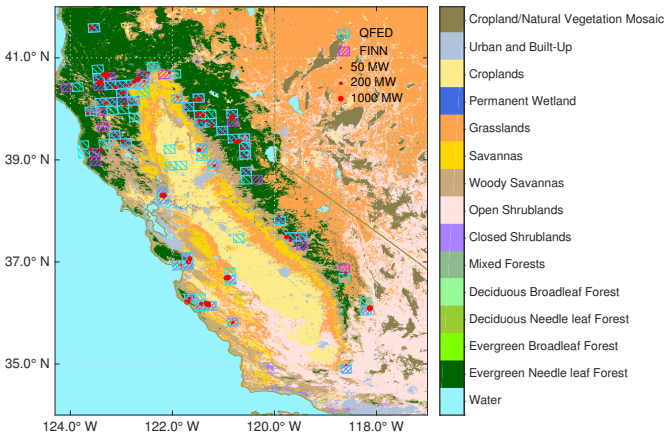

Figure 3.2: Land category types, MODIS fire hot spot detections on 21 and 22 June, 2008, sized by FRP, and 18 km×18 km gridded FINNv1.0 and QFED emission locations.

California presented in Sec. 3.3, the locational differences do affect the results.

3.2.3 WRFDA-Chem inversion system

The aim of data assimilation (DA) is to optimally combine uncertain observations with uncertain model predictions to provide an improved estimate of the state of a system than either gives alone. In Bayesian statistics, the probability distribution of a set of control variables (CV), $\boldsymbol{x} \in \mathcal{R}^n$, conditional on available observations, $\boldsymbol{y}^o$, is proportional to the product of two known distributions,

$$P\left(\boldsymbol{x}|\boldsymbol{y}^o\right) \propto P\left(\boldsymbol{x}\right) P\left(\boldsymbol{y}^o|\boldsymbol{x}\right). \tag{3.1}$$

The first distribution on the right hand side is called the prior, background, or first guess; the second is the likelihood of model-observation mismatch, where here both are assumed to be Gaussian. They are found through the solution of the minimization problem

$$\begin{aligned} \min_{\boldsymbol{x}} \quad J\left(\boldsymbol{x}\right) = & \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{x}_b\right)^{\top} \mathbf{B}^{-1}\left(\boldsymbol{x}-\boldsymbol{x}_b\right) \\ & + \frac{1}{2}\left(G\left(\boldsymbol{x}\right)-\boldsymbol{y}^o\right)^{\top} \mathbf{R}^{-1}\left(G\left(\boldsymbol{x}\right)-\boldsymbol{y}^o\right), \end{aligned} \tag{3.2}$$

where $\boldsymbol{x_b}$ is the vector of prior CVs, $\mathbf{B}$ is the background covariance matrix, and $\mathbf{R}$ is the model-observation error covariance matrix. The nonlinear operator,

$$G(\boldsymbol{x}) = \begin{pmatrix} H_1(\boldsymbol{x}) \\ \vdots \\ H_i(\boldsymbol{x}) \\ \vdots \\ H_N(\boldsymbol{x}) \end{pmatrix}, \tag{3.3}$$

is similar to that applied by Weaver et al. (2005) and Tshimanga et al. (2008), and is composed of the model-observation operators, with each H_i mapping $\boldsymbol{x}$ to observation time i. The measurements at each acquisition time, $y_i^o \in \mathcal{R}^{m_i}$, are expressed independently for N acquisition times by

$$\boldsymbol{y}^o = \left[\boldsymbol{y}_1^{o\top}, \ldots, \boldsymbol{y}_N^{o\top}\right]^\top \in \mathcal{R}^m, \tag{3.4}$$

where $\sum_{i=1}^{N} m_i = m$. The o superscript denotes that $\boldsymbol{y}^o$ are observations.

The cost function in Eq. 3.2 is derived for unbiased Gaussian statistics in both the background errors and model-observation errors. When grid-scale CV uncertainties are greater than 100%, as is often the case for chemical emissions, that assumption allows the posterior to be either positive or negative. While net surface flux rates can be negative when accounting for upward and downward rates together, emission rates are themselves positive. To ensure this, the ratio of modeled (posterior, E_a) to tabulated inventory (prior, E_b) emissions in all grid cells are gathered into a vector, $\boldsymbol{\beta} = e^{\boldsymbol{x}_a}$, such that

$$E_{a,j} = E_{b,j}\beta_j, \tag{3.5}$$

for CV member j. Each β_j is a linear scaling factor, while exponential scaling factors comprise the posterior CV vector, $\boldsymbol{x}_a$. Fletcher and Zupanski (2007) showed that this approach – which was previously utilized in emission inversions by, e.g., Müller and Stavrakou (2005), Elbern et al. (2007), and Henze et al. (2009) – converges toward the median of a multivariate log-normal distribution for $\boldsymbol{\beta}$. Although other emission scaling forms have proven effective (Bergamaschi et al., 2009; Jiang

et al., 2015), we stick with exponential scaling factors here both as a first demonstration, and to be consistent with log-normal statistics for emission rates. $\boldsymbol{x}$ is resolved on the grid scale and across hourly discretized emission rates; the temporal resolution is customizable.

3.2.3.1 Incremental 4D-Var

Here we apply incremental 4D-Var as first introduced by Courtier et al. (1994), utilizing the existing software architecture in WRFDA, and extended to accommodate exponential emission scaling factor CVs. Incremental 4D-Var starts from the assumption that model evaluations of perturbed CVs at observation time k can be expressed by

$$G\left(\boldsymbol{x}+\delta\boldsymbol{x}\right) \approx G\left(\boldsymbol{x}\right) + \mathbf{G}\delta\boldsymbol{x}, \tag{3.6}$$

where $\mathbf{G}$ is the Jacobian. The full matrix is too large to store in memory, but the product $\mathbf{G}\delta\boldsymbol{x}$ is found through the TLM, transforming increments in CV space to perturbations in observation space. With the assumption, Eq. 3.6, the linearized problem is

$$\begin{aligned}\min_{\delta\boldsymbol{x}^k} \quad J\left(\delta\boldsymbol{x}^k\right) = \frac{1}{2}&\left[\delta\boldsymbol{x}^k + \left(\boldsymbol{x}^{k-1}-\boldsymbol{x}_b\right)\right]^\top \mathbf{B}^{-1} \\ &\left[\delta\boldsymbol{x}^k + \left(\boldsymbol{x}^{k-1}-\boldsymbol{x}_b\right)\right] \\ +\frac{1}{2}&\left(\mathbf{G}^{k-1}\delta\boldsymbol{x}^k - \boldsymbol{d}^{o,k-1}\right)^\top \mathbf{R}^{-1} \\ &\left(\mathbf{G}^{k-1}\delta\boldsymbol{x}^k - \boldsymbol{d}^{o,k-1}\right).\end{aligned} \tag{3.7}$$

Each increment, $\delta\boldsymbol{x}^k$, is found in sequential outer loop iterations, where the inner loop solves the quadratic cost function in Eq. 3.7 using linear optimization. k is the number of the current outer loop. The superscript on $\mathbf{G}^{k-1}$ denotes that it is linearized around the state from the previous

iteration, i.e.,

$$\mathbf{G}^{k-1} = \begin{pmatrix} \mathbf{H}_1|_{\boldsymbol{x}^{k-1}} \\ \vdots \\ \mathbf{H}_i|_{\boldsymbol{x}^{k-1}} \\ \vdots \\ \mathbf{H}_N|_{\boldsymbol{x}^{k-1}} \end{pmatrix}. \tag{3.8}$$

$\boldsymbol{d}^{o,k-1}$ is the innovation between observations and model values in the previous iteration:

$$\boldsymbol{d}^{o,k-1} = \boldsymbol{y}^o - G\left(\boldsymbol{x}^{k-1}\right). \tag{3.9}$$

In an emission inversion for a single chemical species, $n = n_x n_y n_t = O\left(10^5 - 10^6\right)$, depending on the domain size and temporal aggregation. Since the number of members in $\mathbf{B}$ is equal to n^2, finding its inverse is computationally infeasible. To circumvent that challenge, Barker et al. (2004) implemented a control variable transform (CVT) through a square root preconditioner, $\mathbf{U}$, in WRFDA. The increment is transformed as $\delta\boldsymbol{x}^k = \mathbf{U}\delta\boldsymbol{v}^k$, where $\mathbf{B} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{U}^\top\mathbf{B}^{-1}\mathbf{U} = \mathbf{I}_n$, and $\mathbf{I}_n \in \mathcal{R}^{n\times n}$ is the identity matrix. The transformed minimization problem is

$$\begin{aligned} \min_{\delta\boldsymbol{v}^k} \quad J\left(\delta\boldsymbol{v}^k\right) = &\frac{1}{2}\left(\delta\boldsymbol{v}^k - \boldsymbol{d}^{b,k-1}\right)^\top \left(\delta\boldsymbol{v}^k - \boldsymbol{d}^{b,k-1}\right) \\ &+\frac{1}{2}\left(\mathbf{G}^{k-1}\mathbf{U}\delta\boldsymbol{v}^k - \boldsymbol{d}^{o,k-1}\right)^\top \mathbf{R}^{-1} \\ &\left(\mathbf{G}^{k-1}\mathbf{U}\delta\boldsymbol{v}^k - \boldsymbol{d}^{o,k-1}\right), \end{aligned} \tag{3.10}$$

where the background departure, summed over all previous outer iterations, is

$$\boldsymbol{d}^{b,k-1} = -\sum_{k_o=1}^{k-1} \delta\boldsymbol{v}^{k_o}. \tag{3.11}$$

In addition to circumventing calculating $\mathbf{B}^{-1}$, the preconditioner reduces the condition number of the problem, speeding up the minimization process.

3.2.3.2 Error covariance

WRFDA-Chem utilizes a very similar CVT as WRFDA, with some modification for the scaling factor control variables. The transform $\delta\boldsymbol{x}^k = \mathbf{U}\delta\boldsymbol{v}^k$ is performed through two separate

operations as $\mathbf{U} = \mathbf{U}_t\mathbf{U}_h$. Although the horizontal transform ($\mathbf{U}_h$) only deals with correlations in the x and y directions, and the temporal transform ($\mathbf{U}_t$) only does so in the temporal dimension, they are both $n \times n$, with sub-matrices along the diagonal of dimension $(n_x n_y) \times (n_x n_y)$ and $(n_t) \times (n_t)$, respectively. The computational overhead of multiplying by either transform is reduced by only handling the non-zero elements. $\mathbf{U}_h$ is carried out using recursive filters (Barker et al., 2004) and the scalar correlation length scale, L_h. $\mathbf{U}_t$ is constructed in a similar fashion as the vertical transform in WRFDA (Barker et al., 2004), except that herein we use all of its eigenmodes. The user specifies the duration of emission scaling factor bins (in minutes), the temporal correlation timescale (L_t, in hours), and the grid-scale relative emission uncertainty, σ_x. WRFDA-Chem converts these selections to a covariance sub-matrix $\mathbf{B}_t = \mathbf{\Sigma C \Sigma} \in \mathcal{R}^{n_t \times n_t}$, where $\mathbf{C}$ is the temporal correlation matrix and $\mathbf{\Sigma} = \sigma_x \mathbf{I}_{n_t}$. $\mathbf{B}_t$ is square, symmetric, and positive-definite. Similar to Saide et al. (2015b), $\mathbf{C}$ is defined using an exponential decay,

$$C_{ij} = e^{-\frac{\Delta t}{L_t}}, \tag{3.12}$$

where Δt is the time elapsed between the beginning of two particular emission steps. The covariance is decomposed into eigenmodes as $\mathbf{B}_t = \mathbf{E}_t \mathbf{\Lambda}_t \mathbf{E}_t^\top$; these are readily calculated, because the dimension of $\mathbf{B}_t$ is the square of the number of emission time steps (e.g., 24 steps for hourly scaling factors in a single day inversion). Throughout the optimization, the temporal transform is carried out through multiplication by

$$\mathbf{U}_t = \begin{bmatrix} \mathbf{E}_t \mathbf{\Lambda}_t^{1/2} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{E}_t \mathbf{\Lambda}_t^{1/2} \end{bmatrix} \tag{3.13}$$

and its transpose.

In general, the prior variances are estimated in the form of multiplicative emission uncertainty in β space (e.g., "factor of 2, 3, 4, etc."), not in the exponential CV ($\boldsymbol{x}$) space. The covariances (off-diagonal terms of $\mathbf{B}$) defined previously are assumed to be applicable in CV space. Transformations between the expectations and covariances of a multivariate log-normal ($\boldsymbol{\beta} \sim \mathcal{LN}\left(\boldsymbol{\mu}_{\beta^0}, \mathbf{B}_{\beta^0}\right)$) and

a Gaussian distribution (i.e., $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}_b, \mathbf{B}_x)$) are derived by, e.g., Halliwell (2015), as

$$\mathbb{E}\left[\boldsymbol{\beta}^0\right]_i = \boldsymbol{\mu}_{\beta^0} = \exp\left(x_{b,i} + \frac{1}{2}\mathbf{B}_{x,ii}\right) \tag{3.14}$$

and

$$\mathbf{B}_{\beta^0,ij} = \exp\left[x_{b,i} + x_{b,j} + \frac{1}{2}\left(\mathbf{B}_{x,ii} + \mathbf{B}_{x,jj}\right)\right]\left(\exp \mathbf{B}_{x,ij} - 1\right), \tag{3.15}$$

respectively, where i and j are general indices coinciding with individual CV members, $\mathbb{E}$ is the expectation operator and exp is the natural exponential function. The subscript β^0 indicates a variable is evaluated in lognormal space in the previous iteration, when $k = 0$, and the subscript x indicates an evaluation in CV space. Since the CVs are normally distributed, $\boldsymbol{x}_b$ is the mean, median, and mode. As Eq. 3.14 shows, the expected value, or mean, of $\boldsymbol{\beta}^0$ is not equal to its median, $\exp x_{b,i}$, the latter being the characteristic we use here. The prior linear scaling factor variances are

$$\sigma_{\beta^0,i} = \exp\left[x_{b,i} + \frac{1}{2}(\sigma_{x_b,i})^2\right]\left[\exp\left(\sigma_{x_b,i}\right)^2 - 1\right]^{\frac{1}{2}}. \tag{3.16}$$

This is identical to the variance transformation between univariate log-normal and Gaussian distributions. With an initial guess of $\sigma_{x_b,i} = 0$, the recursive inverse relation,

$$\begin{aligned}\sigma_{x_b,i} &= \sqrt{\log\left[1 + \frac{\left(\sigma_{\beta^0,i}\right)^2}{\left(\mu_{\beta^0,i}\right)^2}\right]} \\ &= \sqrt{\log\left[1 + \frac{\left(\sigma_{\beta^0,i}\right)^2}{\exp\left(2x_{b,i} + \left(\sigma_{x_b,i}\right)^2\right)}\right]}.\end{aligned} \tag{3.17}$$

converges for reasonable ranges of $\sigma_{\beta^0,i}$, which is the additive uncertainty in β. Earlier emission inversion works (e.g., Elbern et al., 2007) assume that

$$\left(\sigma_{\beta^0,i} + 1\right)^2 \approx \frac{\exp\left(x_{b,i} + \sigma_{x_b,i}\right)}{\exp\left(x_{b,i} - \sigma_{x_b,i}\right)} = \left(\exp \sigma_{x_b,i}\right)^2,$$

which is equivalent to

$$\sigma_{\beta^0,i} + 1 \approx \exp \sigma_{x_b,i}, \tag{3.18}$$

and its inverse

$$\sigma_{x_b,i} \approx \log\left(\sigma_{\beta^0,i} + 1\right). \tag{3.19}$$

$(\sigma_{\beta^0} + 1)$ is the multiplicative error in emissions. For example, $\sigma_{\beta^0} = 2$ gives a factor of three ($\times 3$) uncertainty. In our case $x_{b,i} = 0$ and Eq. 3.18 gives an error in $\sigma_{\beta^0,i} + 1$ less than 3% for $\sigma_{x_b} \in [0, \log(2)]$, but reaches 100% mismatch at $\sigma_{x_b} = \log(4.2)$. Previous works that use Eq. 3.19 with relative emission errors less than $\times 3$ do not warrant corrections. However, utilizing Eq. 3.17 is important for high-resolution inversions of BB sources, where grid-scale uncertainties are probably above that threshold. Sections 3.3.3 and 3.4 include further discussion of emission uncertainty.

The observation-model covariance matrix, $\mathbf{R}$, is assumed diagonal. For each p measurement, the total variance is defined as the sum of observation $\left(\sigma^2_{p,o}\right)$ and model $\left(\sigma^2_{p,m}\right)$ components, following the approach by Guerrette and Henze (2015). $\sigma_{p,m}$ is determined from an ensemble of 156 WRF-Chem model configurations. Each member uses a unique combination of options for PBL mixing, surface layer, LSM, and longwave and shortwave radiation options, as well as includes or excludes microphysics and subgrid cumulus convection. $\sigma_{p,o}$ accounts for instrument precision, representativeness error, and averaging of measurements to the model resolution. We do not use the weighting term previously defined by Guerrette and Henze (2015), because small residuals with low uncertainty do not appear to hinder the inversion process. Refer to that work for more particular details of how $\sigma^2_{p,m}$ and $\sigma^2_{p,o}$ are calculated.

3.2.3.3 Linear optimization

With all of the terms in Eq. 3.10 defined, the linear optimization proceeds as follows. The inner loop seeks the optimal $\delta\boldsymbol{v}^k$, at which point

$$\begin{aligned}\nabla_{\delta\boldsymbol{v}} J = &\left(\delta\boldsymbol{v}^k - \boldsymbol{d}^{b,k-1}\right) \\ &+ \mathbf{U}^\top {\mathbf{G}^{k-1}}^\top \mathbf{R}^{-1}\left(\mathbf{G}^{k-1}\mathbf{U}\delta\boldsymbol{v}^k - \boldsymbol{d}^{o,k-1}\right) \\ &= \mathbf{0}.\end{aligned} \tag{3.20}$$

The action of $\mathbf{G}^{k-1^\top}$ on a vector is calculated with the ADM. Solving for the CVT increment,

$$\begin{aligned}\delta\boldsymbol{v}^k &= \left(\mathbf{I}_n + \mathbf{U}^\top\mathbf{G}^{k-1^\top}\mathbf{R}^{-1}\mathbf{G}^{k-1}\mathbf{U}\right)^{-1}\left(\boldsymbol{d}^{b,k-1} + \mathbf{U}^\top\mathbf{G}^{k-1^\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right)\\ &= -\left[\mathcal{H}_{\delta\boldsymbol{v}}\right]^{-1}\nabla_{\delta\boldsymbol{v}}J|_{\delta\boldsymbol{v}^k=\mathbf{0}},\end{aligned} \tag{3.21}$$

where $\mathcal{H}_{\delta\boldsymbol{v}} = \nabla^2_{\delta\boldsymbol{v}}J$ is the Hessian of Eq. 3.10. $\mathcal{H}_{\delta\boldsymbol{v}}$ and its inverse are too large to store and calculate explicitly. Through an iterative process, the inner loop linear optimization estimates the product of the inverse Hessian with the initial cost function gradient. Finite precision and the problem dimension, n, prevent Eq. 3.20 from being exactly equal to zero. Increasing the number of inner loop iterations to approach such an objective does not necessarily speed up convergence in the nonlinear problem of Eq. 3.2. Large innovations, $\boldsymbol{d}^{o,k}$, may remain after relinearization around the new state $\boldsymbol{x}^k$.

The two linear optimization algorithms available in WRFDA are Conjugate Gradient and the Lanczos recurrence described on p. 493 of Golub and Van Loan (1996). We use Lanczos, which aids the estimation of posterior error as described in Sec. 3.2.3.4. Linear optimization strategies are designed to solve a quadratic problem

$$\begin{aligned}\min_{\hat{\boldsymbol{x}}} \quad & F\left(\hat{\boldsymbol{x}}\right) = \frac{1}{2}\hat{\boldsymbol{x}}^\top\mathbf{A}\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^\top\boldsymbol{b} + c\\ & \mathbf{A}\hat{\boldsymbol{x}} = \boldsymbol{b}.\end{aligned} \tag{3.22}$$

The equivalence of incremental 4D-Var (Eqs. 3.10) and Gauss Newton (GN) to solve Eq. 3.22 is demonstrated in Appendix A; there, we repeat some derivations by Lawless et al. (2005), Gratton et al. (2007), and Tshimanga et al. (2008) using the notation defined herein. The advantage of this equivalence is that any studies pertaining to issues and advances with GN have the potential to inform incremental 4D-Var; we exploit this in Sec. 3.2.3.5 to improve the relinearization behavior for nonlinear CVs.

3.2.3.4 Posterior Error

Posterior uncertainty is a useful measure to diagnose the value of an emission inversion. In a region of linear behavior of the full cost function, Eq. 3.2, and when $\delta\boldsymbol{x}$ is normally distributed,

the posterior covariance, $\mathbf{P}^a$, is equal to the inverse Hessian of Eq. 3.7 (e.g., Thacker, 1989; Fisher and Courtier, 1995):

$$\mathbf{P}^a = \left[\mathcal{H}_{\delta \boldsymbol{x}}\right]^{-1}, \tag{3.23}$$

where

$$\mathcal{H}_{\delta \boldsymbol{x}} = \mathbf{B}^{-1} + \mathbf{G}^{k-1^\top}\mathbf{R}^{-1}\mathbf{G}^{k-1}. \tag{3.24}$$

Combining this with the expression for the Hessian of Eq. 3.10 we used in Eq. 3.21 gives a conversion from the transformed variable space

$$\mathcal{H}_{\delta \boldsymbol{v}} = \mathbf{U}^\top \mathcal{H}_{\delta \boldsymbol{x}} \mathbf{U}. \tag{3.25}$$

Using a Lanczos recurrence to solve the inner loop optimization problem in Eq. 3.10 has the benefit of producing the means to approximate $\left[\mathcal{H}_{\delta \boldsymbol{v}}\right]^{-1}$, which we demonstrate in Appendix A. The final result of that derivation is the posterior error,

$$\begin{aligned} \mathbf{P}^a =& \mathbf{U}\left[\mathcal{H}_{\delta \boldsymbol{v}}\right]^{-1}\mathbf{U}^\top \\ \approx& \mathbf{B} + \sum_{k_i=1}^{l}\left(\lambda_{k_i}^{-1} - 1\right)\left(\mathbf{U}\hat{\boldsymbol{\nu}}_{k_i}\right)\left(\mathbf{U}\hat{\boldsymbol{\nu}}_{k_i}\right)^\top, \end{aligned} \tag{3.26}$$

in terms of the eigenvectors of $\mathcal{H}_{\delta \boldsymbol{v}}$, $\hat{\boldsymbol{\nu}}_{k_i} = \mathbf{Q}_l \hat{\boldsymbol{w}}_{lk_i}$. Each inner iteration, k_i, leading up to the current iteration l of the Lanczos optimization, produces (1) a new *Lanczos vector* in the orthonormal matrix $\mathbf{Q}_l = [\hat{\boldsymbol{q}}_1, .., \hat{\boldsymbol{q}}_l]$ and (2) a new row and column in a tridiagonal matrix $\mathbf{T}_l$, whose ${k_i}^{th}$ eigenpair is $(\lambda_{k_i}; \hat{\boldsymbol{w}}_{lk_i})$. $\mathbf{P}^a$ is a low-rank update to $\mathbf{B}$, because $l << n$ due to the wall-clock requirements of running the TLM and ADM once per iteration. Equation 3.26 is consistent with earlier publications (Fisher and Courtier, 1995; Meirink et al., 2008).

3.2.3.5 Damped Gauss Newton

Each CV increment, $\delta \boldsymbol{x}^k$, must be small enough to keep the error associated with the tangent linear assumption, Eq. 3.6, below some threshold. However, the nonlinearity of the log-normal prior

emission errors contributes to failures in that respect. For demonstration, we consider the treatment of $\beta \in \boldsymbol{\beta}$ and $x \in \boldsymbol{x}$ associated with a single grid cell. At the end of an outer loop, x is updated, and β is relinearized using

$$\beta^k = e^{x^{k-1}+\delta x^k} = \beta^{k-1} e^{\delta x^k}. \tag{3.27}$$

Thus, the increment in β is

$$\delta\beta^k = \beta^k - \beta^{k-1} = \beta^{k-1}\left(e^{\delta x^k} - 1\right), \tag{3.28}$$

which reveals the nonlinear nature of the emission increment. This contrasts with the TLM version of the transform in Eq. 3.5, which states

$$\delta\beta' = e^{x^{k-1}}\delta x = \beta^{k-1}\delta x'. \tag{3.29}$$

The ratio of $\delta\beta^k/\delta\beta'$ gives the multiplicative error in the tangent linear assumption during relinearization:

$$\epsilon_{TL} = \frac{e^{\delta x^k} - 1}{\delta x^k}. \tag{3.30}$$

Around $\delta x^k = 0$, the tangent linear relationship very closely matches the nonlinear equation, giving $\epsilon_{TL} \approx 1$. For $\delta x^k > 0$, ϵ_{TL} grows nearly exponentially toward ∞, reaching $\times 2$ at $\delta x^k \approx 1.26$. When $\delta x^k < 0$, ϵ_{TL} shrinks asymptotically toward zero, reaching $\times 0.5$ at $\delta x^k \approx -1.59$. As ϵ_{TL} is farther from unity, it is more likely that the linear optimization will generate $J\left(\boldsymbol{x}^k\right) > J\left(\boldsymbol{x}^{k-1}\right)$. Not only do we never want that to happen, but we would prefer to advance toward a more optimal solution as quickly as possible.

Violation of the TL assumption and potential solutions are discussed in several DA works. The prevailing strategy in chemical 4D-Var is to apply a non-incremental nonlinear optimization strategy (e.g., Henze et al., 2009; Bergamaschi et al., 2009), eliminating the inner-outer loop structure. Implementing this approach in WRFDA with posterior error estimation would be a considerable additional effort. The use of the tangent linear model in the inner loop also presents computational advantages for dual resolution multi-incremental 4D-Var (e.g., Zhang et al., 2014b).

Alternatively, Gratton et al. (2013) discuss application of GN in a trust region framework, which has the limitation that a portion of the computationally expensive outer loop increments will be rejected. Some authors have successfully applied the Levenberg-Marquardt algorithm in EnKF (e.g., Chen and Oliver, 2013; Mandel et al., 2016) by adding a regularization term to the cost function.

A simpler approach yet is damped GN (DGN), which changes the inner loop increment in Eq. 3.21 to

$$\delta \boldsymbol{v}^k = -\,\eta^k \left[\mathcal{H}_{\delta \boldsymbol{v}}\right]^{-1} \nabla_{\delta \boldsymbol{v}} J|_{\delta \boldsymbol{v}^k = \boldsymbol{0}}, \tag{3.31}$$

and uses a line search to find an optimal scalar $\eta^k \in (0, 1]$ at the completion of each outer loop iteration (Kelley, 1999). DGN is based on the Armijo rule, which states that the increment found by GN points toward a direction of lower J; if the step size terminus is outside the linear behavior of the model, decrease the step size. WRFDA-Chem uses a non-optimal variant we call heuristic DGN, that requires user intervention to determine η^k. Results with a simplified test problem in MATLAB indicate that the resultant CV's near the optimum are nearly identical either with the line search or heuristic damping. However, heuristic DGN likely increases the number of outer iterations required to converge, and motivates implementing the line search in future work. The same MATLAB tests showed that applying a range of damping coefficient values before the Lanczos process has no impact on the estimated $\mathcal{H}_{\delta \boldsymbol{v}}^{-1}$.

The heuristics to determine η^k are a function of the prior covariance. As the uncertainty increases, η^k should be smaller, because the initial gradient and resulting increments will be larger in magnitude. Additionally, η^k should increase in each subsequent outer loop iteration as the nonlinear optimum is approached, since the diminishing increment magnitude will eventually satisfy the tangent linear assumption. We found that a prior multiplicative emission uncertainty of $\times 3.8$, coinciding with CV uncertainty of $\sigma_x = 1.099$, requires $\eta^0 = 0.4$ in the first outer loop iteration. η^0 should be adjusted in inverse proportion to σ_x. Presumably there is some lower limit of σ_x where no damping is required. In WRFDA-Chem, the damping ramps linearly back to 1 in the final outer loop.

3.3 ARCTAS-CARB Case Study

3.3.1 Inversion setup

From late May until 20 June 2008, the southwest U.S. experienced a very dry period with little to no cloud cover appearing in MODIS true color imagery, and no recorded rainfall for most of California. On 21 June, the Aqua and Terra satellites recorded cloud cover for much of Northern California, south of San Francisco, and along the Sierra Nevada mountain range, and there were wide-spread lightning strikes over night. As is shown in Fig. 3.1, there was a spike in fire detections during the night between 21 and 22 June. Thus, from the morning through evening of 22 June, California experienced a transient fire initiation event. The wild fires burned well into July, exacerbating poor air quality throughout the state. The 22 June flight of ARCTAS-CARB disembarked from Los Angeles, swept out over the ocean, flew directly through smoke from forest fires in Northern California, then returned down the coastline. That flight encountered anthropogenic sources of BC in the morning, and BB sources for the remainder after returning to land. The 24 June flight passed back and forth in the downwind region between Los Angeles and San Diego, measuring the outflow from those cities and the transportation lines between them. A third flight on 26 June flew in the free troposphere from Los Angeles, north over the fires, and exited the model domain to the east.

We use WRFDA-Chem 4D-Var to constrain BB and anthropogenic aerosols on three days during ARCTAS-CARB using aircraft and IMPROVE surface observations. We utilize aircraft measurements of absorbing carbonaceous aerosol at 10 s intervals from the single particle soot photometer (SP2) on 22, 24, and 26 June (Sahu et al., 2012). For this study, we assume equivalency between the SP2 measurement and modeled BC, and re-average to the 90 s model time step using the revision 3 product, a process described in Guerrette and Henze (2015). We also use 24-hour average surface observations of light absorbing carbon (LAC) on 23 and 26 June (Malm et al., 1994), assuming an equivalence with modeled BC, and ignoring the 7% high bias relative to the SP2 found by Yelverton et al. (2014). All treatments of observations are identical to those described

in Guerrette and Henze (2015), including an analysis of model-observation BC mismatch that feeds into the inverse modeling study.

Using measurements from 22, 23, and 24 June, the 4D-Var system constrains anthropogenic and BB sources simultaneously. Data collected between 07:00:00-16:00:00 LT on 22 June is used in an inversion from 22 June, 00:00:00 UTC to 23 June, 00:00:00 UTC, during which time WRF-Chem is run freely, without nudging. The emission scaling factors for this 24-hour time period for both source types are applied to subsequent days from 23-26 June in a cross validation experiment. The 24 and 26 June aircraft and 23 and 26 June surface observations are used to analyze the utility of observationally constrained scaling factors found on one day to fix source errors on subsequent days. The 23 and 24 June surface and aircraft data is used in a 48-hour inversion from 23 June, 00:00:00 UTC to 25 June, 00:00:00 UTC, also without nudging. Cross validation is performed for these source estimates using 26 June surface and aircraft data.

Through preliminary testing, we found that horizontal correlation length scales on the order of the grid spacing provides the lowest posterior cost function. For both time periods, this length scale is set to twice the grid scale, $L_h = 36$ km. The emission scaling factors are aggregated in each hour, which coincides with the emission file reading interval for both source types. The correlation scale is set to $L_t = 4$ h, following Saide et al. (2015b). In addition to spreading error information across adjacent grid cells, the correlation scales reduce the effective number of CVs. Through sensitivity tests where we considered the smoothness of the posterior and the stationary posterior cost function value, and after consulting published values for regional emission uncertainties (see Sec. 3.1) in different global settings, we use a grid-scale BB uncertainty of $\times 3.8$. The BB uncertainty might also be approximated from the ratio of prior domain-wide total emissions between FINNv1.0 and QFED, which is given in Table 3.1 as $\times 3.5$. If the median emission strength lies in the middle of QFED and FINNv1.0, then the prior domain-wide relative uncertainty is $\times\sqrt{3.5} = \times 1.8$. The uncertainty would then need to be inflated further to account for spatial and temporal disaggregation and the possibility that grid-scale sources from the two inventories do not bound the true value. The prior anthropogenic grid-scale uncertainty is set to $\times 2$, which is within the reasonable bounds discussed

in Sec. 3.1.

Table 3.1: Total BB emissions for EA's and domain-wide during 22 and 23/24 June inversions (averaged for 24 hour period). Absolute units are in Mg. Note, the differences (Δ) may not sum due to rounding.

		FINN_STD			QFED_STD			$\frac{\Sigma E_{\mathrm{QFED}}}{\Sigma E_{\mathrm{FINN}}}$	
		ΣE_b	ΣE_a	Δ	ΣE_b	ΣE_a	Δ	b	a
22 June	EA1	14	4	-10	82	26	-55	×5.8	×6.4
	EA2	6	30	+24	9	15	+6	×1.5	×0.5
	EA3	6	4	-2	29	7	-22	×4.5	×1.6
	EA4	18	22	+4	52	83	+31	×2.8	×3.8
	DOMAIN	59	83	+34	209	171	-38	×3.5	×2.1
23+24 June	EA1	20	5	-15	70	12	-58	×3.5	×2.5
	EA2	28	11	-16	96	29	-67	×3.5	×2.6
	EA3	17	12	-5	37	20	-17	×2.2	×1.7
	EA4	32	108	+77	107	107	0	×3.4	×1.0
	DOMAIN	138	249	+111	471	354	-117	×3.4	×1.4

In addition to these standard settings, several sensitivity scenarios are used to gauge the sensitivity of the posteriors during two time periods to alternative inversion settings. The full set of scenarios are summarized in Table 3.2, and are as follows. FINNv1.0 is used as the default BB inventory in a scenario called FINN_STD for both inversion periods. QFED_STD uses the QFEDv2.4r8 BB inventory. Both FINN_L18 and QFED_L18 use $L_h = 18$ km. FINN_V1.5 utilizes the FINNv1.5 BB inventory. For the 23/24 June inversion, we show results for both QFED_STD and FINN_STD, the latter of which includes variations where either surface or aircraft observations are excluded. The number of aircraft observations is $N_{\mathrm{obs}} = 241$ on 22 June and $N_{\mathrm{obs}} = 302$ on 24 June. There were $N_{\mathrm{obs}} = 35$ active surface sites on 23 June, 13 of them within California. We use six outer iterations consisting of 10 inner iterations each. Given the number of inner iterations used, and the wall-time of the tangent linear plus the adjoint (10× the nonlinear model), the cost of incremental 4D-Var is approximately 600 × that of a single forward simulation, which is much cheaper than using finite difference methods to approximate derivatives instead of the linearized models when$n \sim 10^5$.

Table 3.2: Emission inversion scenarios.

	Scenario	BB Inventory	L_h	Obs. Used (day)
22 June	FINN_STD	FINNv1.0	36 km	ARCTAS-CARB (22)
	FINN_L18	FINNv1.0	18 km	ARCTAS-CARB (22)
	QFED_STD	QFEDv2.4r8	36 km	ARCTAS-CARB (22)
	QFED_L18	QFEDv2.4r8	18 km	ARCTAS-CARB (22)
	FINN_V1.5	FINNv1.5	36 km	ARCTAS-CARB (22)
23/24 June	FINN_STD	FINNv1.0	36 km	IMPROVE (23) & ARCTAS-CARB (24)
	QFED_STD	QFEDv2.4r8	36 km	IMPROVE (23) & ARCTAS-CARB (24)
	ACFT	FINNv1.0	36 km	ARCTAS-CARB (24)
	SURF	FINNv1.0	36 km	IMPROVE (23)

3.3.2 Posterior model performance

The convergence properties of the 22 and 23/24 June inversion scenarios are shown in the outer loop cost function progression in Fig. 3.3. All of the 22 June scenarios led to comparable cost function values at numerical convergence, as shown in Fig. 3.3. The gradient norms are also reduced by nearly two orders of magnitude in all cases. The χ^2 criteria states that the posterior cost function should be equal to $\frac{1}{2}N_{\text{OBS}}$. In all of the scenarios, J converges to approximately N_{OBS}, indicating that a portion of the model errors are not fully spanned by prior emission errors. For the 23/24 June inversion, QFED_STD reaches a lower cost function value, and both scenarios achieve similar χ^2 values as the 22 June cases. Scrutinizing other sources of error (e.g., initial and boundary conditions for BC and meteorological variables, transport, BB plumerise, and model discretization) either independent from source strengths or simultaneously in the inversion framework should elicit further cost function reductions.

The non-emission sources of error for 22 June are evident in the time series in Fig. 3.4. The posterior is within the combined model/observation uncertainty (see Sec. 3.2.3.2) much more often than the prior. The only time during the inversion when the forecast degrades is for an observed peak at 22 June, 08:00 LT. Model uncertainty is higher in locations where the prior concentration

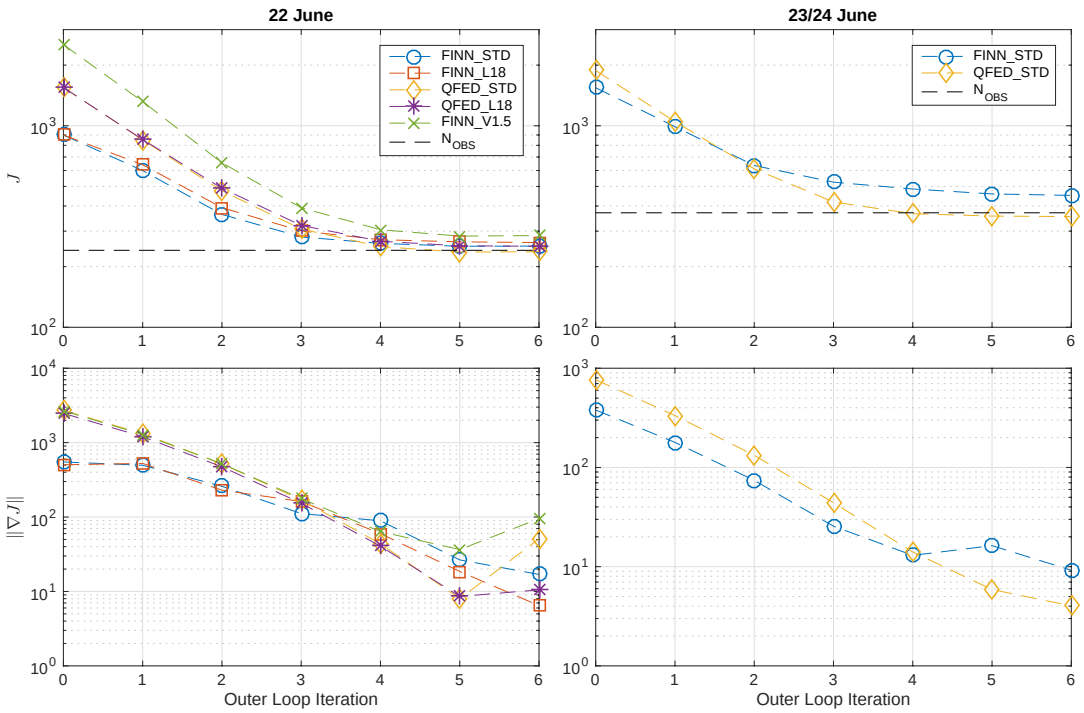

Figure 3.3: Outer loop cost function and gradient norm evaluations for the June 22 (left column) and 23/24 June (right column) inversions.

is higher, due to variability in the configuration ensemble boundary layer heights (Guerrette and Henze, 2015). This high prior uncertainty in **R** allows the stronger constraint at 08:30 LT to dominate the morning anthropogenic emissions, since this flight portion was confined to Los Angeles. In the afternoon, when the DC-8 passed over the wildfires, an increase in posterior emissions captures several of the observed BC peaks. The posterior is able to match the high-resolution variability of the observations at 13:30 LT, which may support the validity of the temporal averaging scheme.

The R^2 coefficients and slopes for linear fits between the prior and posterior and both aircraft and surface observations are summarized in Tables 3.3 and 3.4. Those results include cross-validation data on non-inversion days, which is discussed in Sec. 3.3.5. For both inversion periods, there are considerable model performance improvements for observations that are used in the inversion. FINN_STD improves R^2 from 0.11 to 0.82 and slope from 0.26 to 0.8 on 22 June. QFED_STD improves R^2 from 0.03 to 0.73 and slope from 0.34 to 0.71. Similar improvements occur for 23 June surface observations during the 23/24 June inversion. The posterior match to 24 June aircraft observations is improved, but not nearly as much as the other two data sets. The 22 June inversion results are also shown in the first row of Fig. 3.5, where the progression of the fit parameters is shown for the multiple scenarios. While all scenarios show similar improvements, the FINN_STD and QFED_STD results indicate the posteriors are still underpredicting many low and high concentrations. Overprediction seems to be less of a problem. A similar phenomenon occurs for the 24 June observations in Fig. 3.6 in the inversion that uses both surface and aircraft observations. On both 22 and 24 June, the remaining low bias is either due to large prior observation and model error (diagonal of **R**) or due to the prior errors not being sensitive to emission increments.

3.3.3 Posterior emissions

Figure 3.8 shows the prior and posterior BB emissions for FINN_STD and QFED_STD during both simulation periods. In that figure there are several outlined emission areas (EAs); each EA was chosen to identify regions where a subset of the grid-scale analysis increment ($\delta\boldsymbol{x}_{\mathrm{EAX}} \subset \delta\boldsymbol{x}$) from both prior inventories is of similar sign. The coordinates of the EAs are listed in Table 3.5.

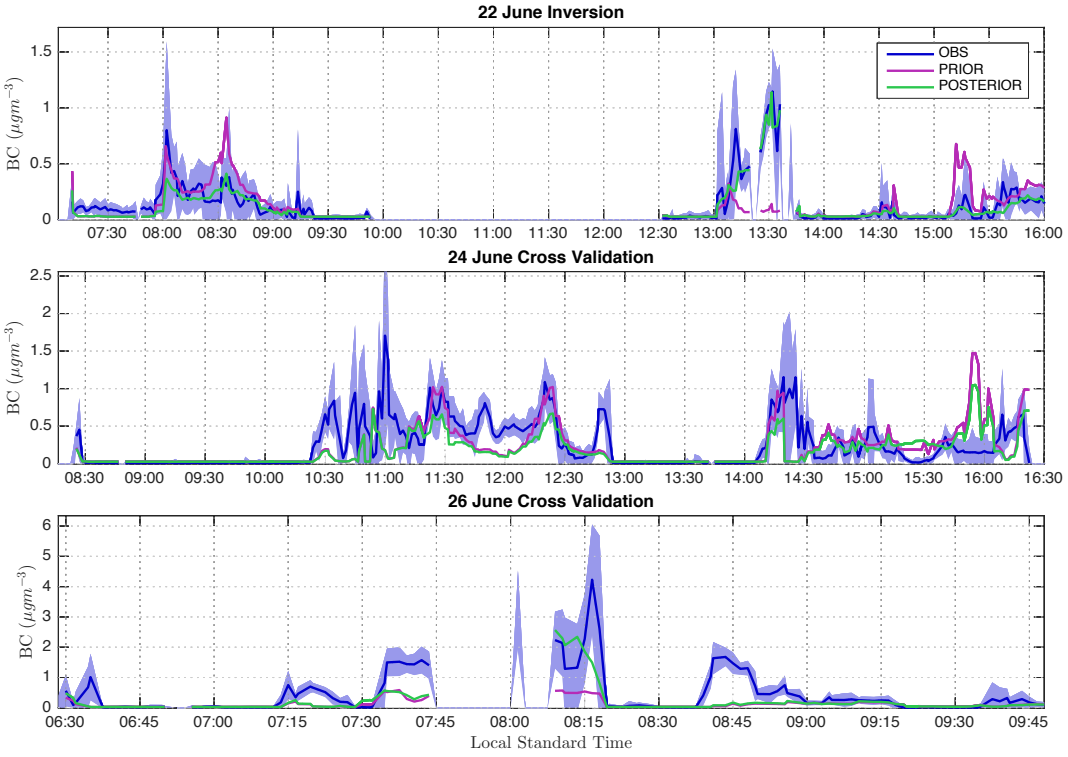

Figure 3.4: Temporal variation of observed, prior, and posterior BC concentrations during ARCTAS-CARB. The model values are obtained with the FINN_STD inversion scenario. The shaded area encompasses 2 standard deviations around the observations, which includes both model and observation uncertainty.

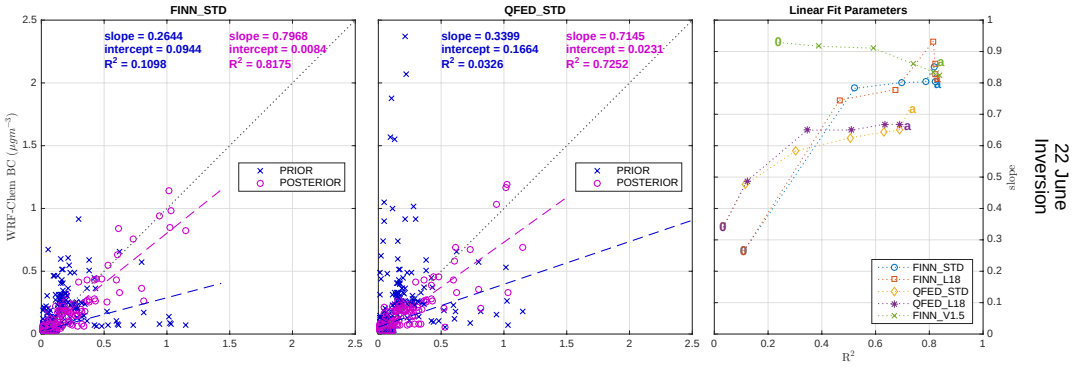

Figure 3.5: Prior and posterior model versus 22 June ARCTAS-CARB observations for the 22 June inversion. The left two plots are for FINN_STD and QFED_STD. The plot on the right shows the progression of slope and R^2 from the prior, "0", to the posterior, "a", for similar linear regressions in all scenarios.

Table 3.3: Aircraft observation linear regression characteristics for the prior (background, b) and posterior (analysis, a).

	Obs. Date →	22 June, $N_{obs} = 241$				24 June, $N_{obs} = 301$				26 June, $N_{obs} = 117$			
	Inversion Scenario	R^2		slope		R^2		slope		R^2		slope	
	↓	b	a	b	a	b	a	b	a	b	a	b	a
22 June	FINN_STD	0.11	**0.82**	0.26	**0.80**	0.18	0.15	0.38	*0.25*	0.56	0.52	0.15	**0.49**
	QFED_STD	0.03	**0.73**	0.34	**0.71**	0.15	0.23	0.43	0.37	0.59	0.53	0.39	0.43
23/24 June	FINN_STD	-	-	-	-	0.17	**0.52**	0.35	**0.56**	0.59	*0.16*	0.15	0.11
	QFED_STD	-	-	-	-	0.11	**0.52**	0.36	**0.55**	0.63	*0.44*	0.41	*0.15*
	ACFT	-	-	-	-	0.17	**0.53**	0.35	**0.57**	0.59	*0.29*	0.15	0.08
	SURF	-	-	-	-	0.17	0.17	0.35	0.40	0.59	*0.13*	0.15	0.17

distinct improvement; *distinct degradation*; cross validation

The two inversions do not reach identical total posterior BC emissions, but they do converge in certain aspects. Table 3.1 gives the emission subtotals for the EAs. During both inversions, each EA has emission increments of the same sign for both scenarios. Therefore, while domain-wide sources seem to be bounded by the two priors (as evidenced by their convergence), the same might not be true within the individual EAs. EA3, which accounts for the smallest average posterior total, is the only region where the magnitude of the log-ratio between QFED and FINN is smaller in the posterior on 22 June. The ratio is reduced in EA2, but there the FINN posterior is $\times 2$ larger than that for QFED. On 23/24 June, the two scenarios have less posterior spread in all of the EAs. Although Table 3.1 indicates large changes in source strengths across the EAs, Figure 3.7 reveals that a majority of the absolute emission increment (posterior minus prior) in both FINN_STD and QFED_STD arose in only a few grid cells, often where the prior has the largest magnitude. The linear scaling factor pattern is similar between the two scenarios, with those for QFED_STD shifted toward decreases due to the high prior bias.

The temporal distribution of prior and posterior BB emissions within the four EAs are shown in Fig. 3.9 across all inversion scenarios on 22 June. The FINNv1.5 prior is an extreme outlier on the local afternoon of 21 June for EA1, EA2, and EA4. The same is true all day on 22 June for EA2, where the posteriors from other scenarios adjust toward the FINNv1.5 prior. Meanwhile, in

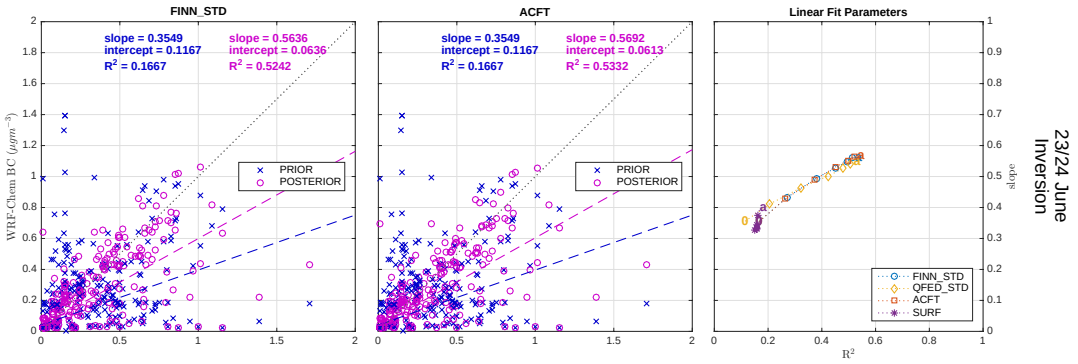

Figure 3.6: Prior and posterior model versus 24 June ARCTAS-CARB observations for the 23/24 June FINN_STD inversion. The left plot uses both IMPROVE (23 June) and ARCTAS-CARB observations in the inversion. The middle plot uses only ARCTAS-CARB. The plot on the right shows the progression of slope and R^2 from the prior, "0", to the posterior, "a", for similar linear regressions.

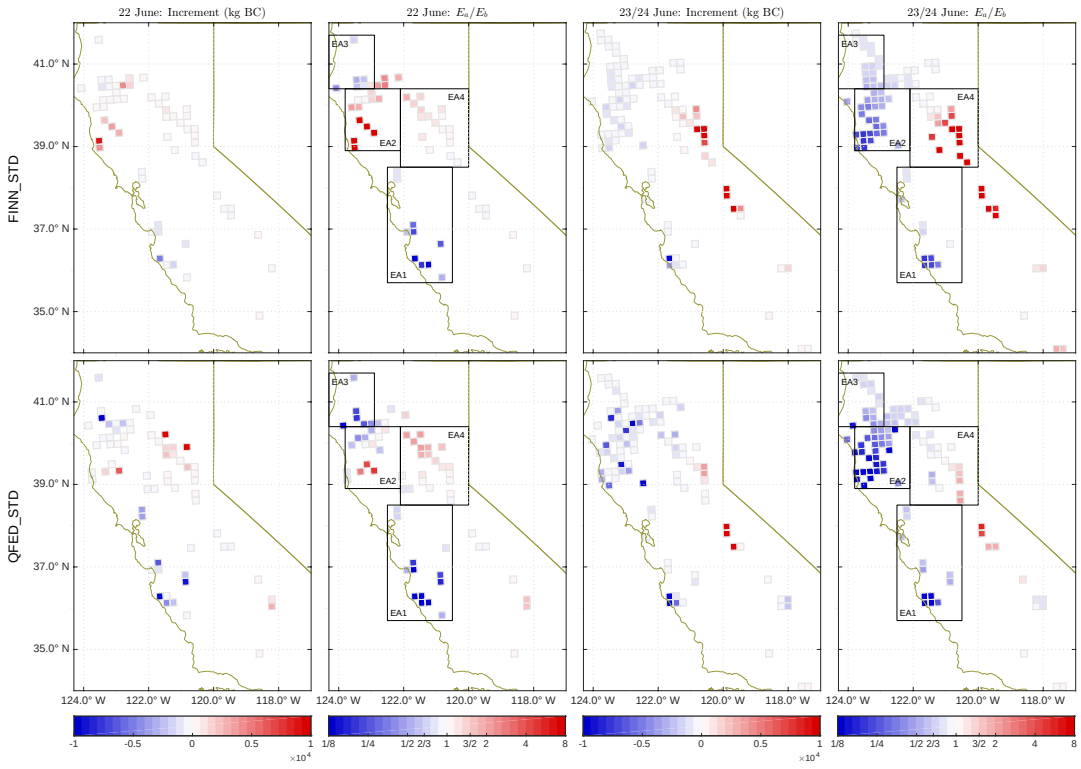

Figure 3.7: BB analysis increment (posterior minus prior) per 24 hours and posterior linear scaling factor (β) for the two primary BB scenarios on 22 June 00Z-23Z and 23 June, 00Z - 24 June 23Z. EA1-4 are outlined with black boxes. [NOTE on Figures 2, 7 and 8: we are waiting on results from QFED_STD with IMPROVE obs included.]

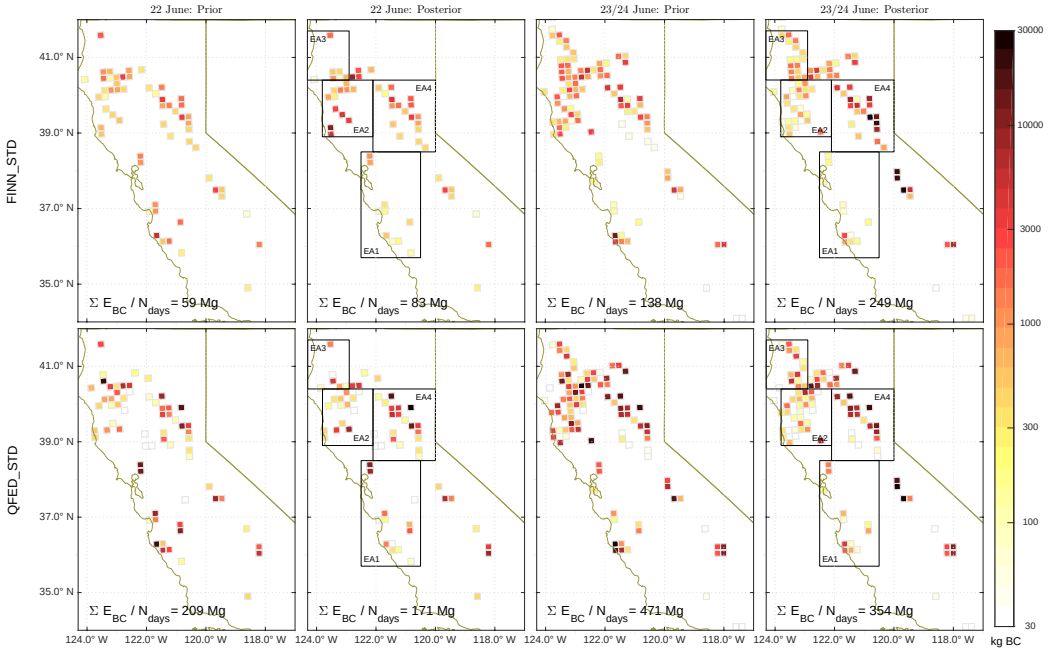

Figure 3.8: Prior and posterior grid-scale BB emissions of BC per 24 hours for FINN_STD and QFED_STD on 22 June, 00Z-23Z and 23 June, 00Z - 24 June 23Z. All emissions are expressed for a 24 h average. EA1-4 are outlined with black boxes.[NOTE on Figures 2, 7 and 8: we are waiting on results from QFED_STD with IMPROVE obs included.]

Table 3.4: Surface observation linear regression characteristics for the prior (background, b) and posterior (analysis, a).

	Obs. Date →	23 June, $N_{obs} = 35$				26 June, $N_{obs} = 36$			
	Inversion Scenario	R^2		slope		R^2		slope	
	↓	b	a	b	a	b	a	b	a
22 June	FINN_STD	0.06	0.04	0.26	0.21	0.03	0.05	0.10	0.13
	QFED_STD	0.16	0.14	0.44	0.41	0.10	0.11	0.20	0.21
23/24 June	FINN_STD	0.04	**0.75**	0.25	**1.04**	0.03	**0.28**	0.10	**0.28**
	QFED_STD	0.09	**0.74**	0.39	**1.01**	0.09	0.15	0.20	0.16
	ACFT	0.04	0.05	0.25	0.27	0.03	0.03	0.10	0.09
	SURF	0.04	**0.74**	0.25	**1.02**	0.03	**0.35**	0.10	**0.35**

distinct improvement; *distinct degradation*; cross validation

other times when FINNv1.5 appears to converge toward the posteriors found using the other two priors, the prior uncertainty of $\times 3.8$ is too restrictive to allow full convergence, since the priors differ by $\times 10$. EA1 is characterized by decreases for all scenarios at all times. EA2, EA3, and EA4 exhibit early morning peaks between 03:00 and 06:00 LT that were not captured in the prior. In separate sensitivity tests, these peaks only appear when $L_t > 1$ h, and become more prevalent as L_t is increased. Saide et al. (2015b) attributed similar behavior in posterior estimates of the 2013 Rim Fire to persistent large scale burning. Zhang et al. (2012) found similar, less pronounced bimodal behavior for all of North America, which could be more noticeable in a regional inversion. Another possibility on 22 June 2008 is that the early morning burning is caused by the transient fire initiation event, which would explain the ramping of emissions for the QFED and FINNv1.5 posteriors in EA2. For both QFED and FINNv1.0, reducing the correlation length to $L_h = 18$ km reduces the analysis increment in all EAs. This is especially apparent in EA4 for FINN_L18, where the increment is negligible.

The differing diurnal patterns in EA2 across scenarios could be attributed to variation in plume heights, QFED regridding errors, and the regularization term of the cost function. The observations most sensitive to EA2 sources were captured within or very near fire plumes. Plume heights are calculated hourly in an online 1D vertical mixing scheme in WRF-Chem (Freitas et al.,

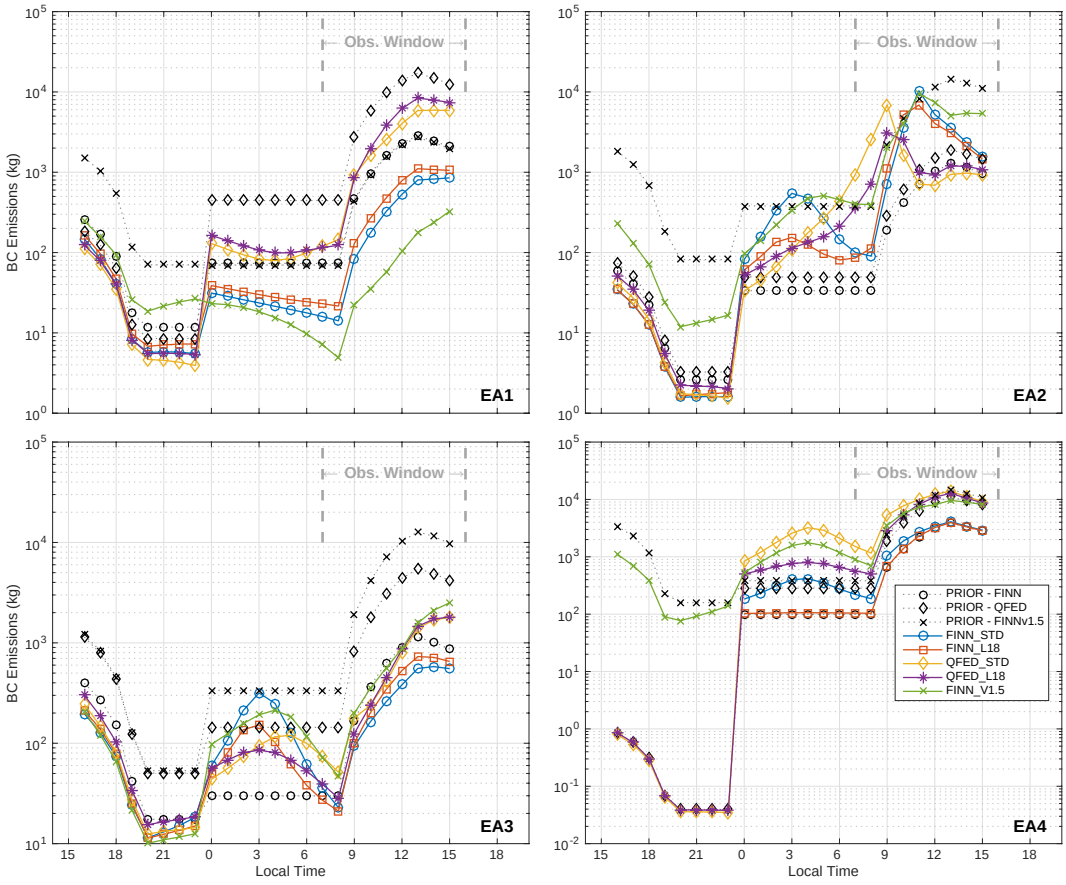

Figure 3.9: Hourly prior and posterior BB diurnal emission patterns for the four EAs and all inversion scenarios for 22 June, 00Z-23Z, with the time shown in LT. Note that FINNv1.0 did not have any fires in EA4 on 21 June.

Table 3.5: Emission area coordinates. EA1-4 are used for BB totals and EA5-9 are used for anthropogenic totals.

	LON_{min}	LON_{max}	LAT_{min}	LAT_{max}
EA1	122.5°W	120.5°W	35.7°N	38.5°N
EA2	123.8°W	122.1°W	38.9°N	40.4°N
EA3	124.3°W	122.9°W	40.4°N	41.7°N
EA4	122.1°W	120.0°W	38.5°N	40.4°N
EA5	117.8°W	116.9°W	32.1°N	33.4°N
EA6	121.0°W	117.8°W	33.4°N	34.6°N
EA7	123.0°W	121.0°W	36.6°N	38.8°N
EA8	120.6°W	118.6°W	35.2°N	37.0°N
EA9	118.0°W	116.5°W	34.0°N	36.0°N
EA10	116.9°W	115.0°W	32.1°N	33.4°N

2007, 2010; Grell et al., 2011), which depends strongly on burned areas. With FINN, the areas are provided for each fire independently, while for QFED the areas use a default value of 0.25 km^2 per fire. In both cases, the maximum area burned per grid cell per day is 2 km^2. The regridding error discussed in Sec. 3.2.2 introduces fire locational errors, especially in EA2. A small error in vertical or horizontal mapping of a discrete point source on the model grid could hinder the optimization in distinguishing it from others. The uniform relative uncertainty in the prior inhibits consolidation of multiple posteriors when the prior spread is heterogeneous and sometimes very large. Quantifying the heterogeneity of uncertainty could contribute to posterior agreement between inversions using different priors, as well as to reducing the cost function.

The spread of local emissions provide some sense of that heterogeneity. Each EA covers a region approximately the size of a grid box in a global simulation with a chemical transport model. Due to the nature of variance aggregation, uncertainty grows as the grid scale gets smaller. In individual EAs, the spread between FINNv1.0 and QFED priors is ×2-×6 for both hourly (Fig. 3.9) and daily (Table 3.1) strength on 22 June. If the median emission strength lies in the middle, then a proxy for prior EA relative uncertainty is $\times\sqrt{2} - \times\sqrt{6} = \times 1.4 - \times 2.4$. Since the two inventories use identical diurnal patterns, the hourly estimate is missing information about uncertainties in daily emission timing. Using the posterior spread in a similar way gives approximate EA uncertainties of

$\times\sqrt{3} - \times\sqrt{10} = \times 1.7 - \times 3.2$ on hourly scales and $\times\sqrt{2} - \times\sqrt{7} = \times 1.4 - \times 2.6$ on daily scales. This posterior estimate accounts for contributions in the prior definitions, including regridding, plume rise, and diurnal patterns. These ranges provide much more detail estimates than simply taking the domain-wide ratio of total emissions for the campaign period. However, the spread is itself missing information about uncertainty that could be found through carrying out similar inversions across an ensemble of model configurations and meteorological initial and boundary conditions (e.g., Lauvaux et al., 2016), or by comparing many more inventory priors (e.g., Zhang et al., 2012) and posteriors. All this is to say that the BB inventories used in this study are not provided with analytical estimates of uncertainty, and a lack of information for deriving such values at hourly grid-scales is a topic for future research.

Figure 3.11 shows the total prior and posterior anthropogenic emissions and Fig. 3.10 displays the analysis increment and linear scaling factor for FINN_STD on 22 June and separately on 23-24 June. The only difference in QFED_STD, not shown here, is that anthropogenic scaling factors are shifted in the negative direction in the posterior, likely due to the higher bias in that BB prior. The increments found in a new set of EAs are presented in Table 3.6.

Table 3.6: Total anthropogenic emissions for EA's and domain-wide during 22 and 23/24 June inversions (averaged for 24 hour period). The posterior for 23/24 June is from an inversion using both the IMPROVE and ARACTAS-CARB observations. Results shown are for the FINN_STD scenario. Absolute units are in Mg. Note, the differences (Δ) may not sum due to rounding.

	22 June			23/24 June			$\frac{\Sigma E_{23+24\text{ June}}}{\Sigma E_{22\text{ June}}}$	
	ΣE_b	ΣE_a	Δ	ΣE_b	ΣE_a	Δ	b	a
EA5	5	5	0	7	3	-3	$\times 1.4$	$\times 0.7$
EA6	12	8	-4	17	9	-8	$\times 1.4$	$\times 1.2$
EA7	10	6	-5	16	8	-8	$\times 1.6$	$\times 1.5$
EA8	3	2	-1	5	25	+20	$\times 1.6$	$\times 9.9$
EA9	5	4	-1	6	11	+4	$\times 1.3$	$\times 2.7$
EA10	2	2	0	3	8	+5	$\times 1.4$	$\times 3.7$
DOMAIN	81	68	-13	114	123	+9	$\times 1.4$	$\times 1.8$

The 23 and 24 June observations provide much more detailed information about anthro-

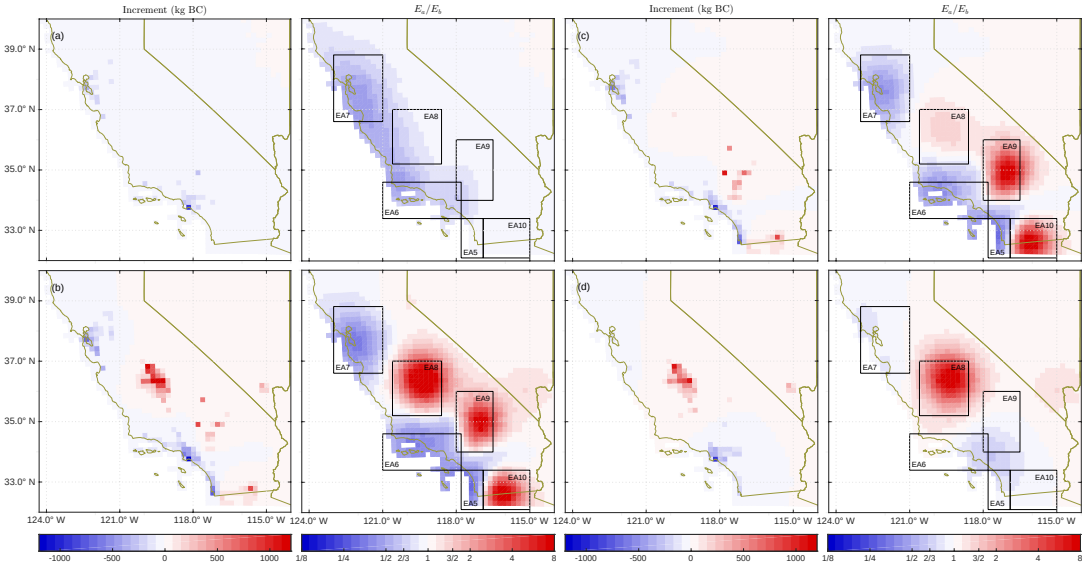

Figure 3.10: Anthropogenic analysis increment (posterior minus prior) per 24 hours and posterior linear scaling factor (β) for the (a) FINN_STD (22), (b) FINN_STD (23/24), (c) ACFT, and (d) SURF inversion scenarios. EA5-9 are outlined with black boxes in the scaling factor plots.

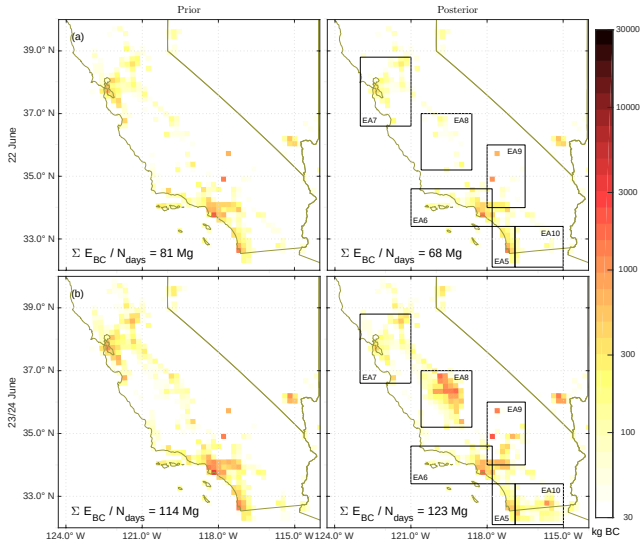

Figure 3.11: Prior and posterior grid-scale anthropogenic emissions of BC per 24 hours for FINN_STD on 22 June, 00Z-23Z (top row) and 23 June, 00Z to 24 June, 23Z. EA5-10 are outlined with black boxes.

pogenic sources. The analysis increment reveals potentially misrepresented city-level emissions in the NEI05 prior. Posterior BC near Barstow, Victorville/Hesperia, Fresno, Edwards Air Force Base, and El Centro/Calexico are increased, while sources near the three coastal cities are decreased. Since Barstow is a crossroads for the BNSF and the Union Pacific railroads, and since Fresno, Victorville/Hesperia, and El Centro/Calexico lie at switching locations for major rail lines, it could be speculated that the prior is missing diesel rail sources of BC. Another possibility is that low bias fire emissions north of Fresno are responsible for the prior underpredictions of 23 June surface concentration measurements exceeding 2 $\mu g\,m^{-3}$ (see Fig. 5 of Guerrette and Henze, 2015). This is corroborated by the posterior BB emissions being scaled up near Fresno on 23 and 24 June, and by the much smaller model bias for IMPROVE on 22 June before the fires started.

There are also small negative increments near Los Angeles (EA6) and San Francisco (EA7) during both the 22 June and 23/24 June inversions, which are likely attributable to on-road mobile sources. These results are consistent with model bias in surface and aircraft observations on 20 June near both of those cities (Guerrette and Henze, 2015). McDonald et al. (2015) found a decreasing trend in ambient measurements of BC and in a fuel-based bottom-up inventory for both Los Angeles and San Francisco from 1990 to 2010 that might not be captured for the 2008 model year by the snapshot in NEI05. Using a similar fuel-based approach, Kim et al. (2016) derived 2010 CO emissions in the South Coast Air Basin surrounding Los Angeles that are $\times\frac{1}{2}$ the magnitude of NEI05. On-road and other mobile sources make up 36% and 62% of that difference, respectively, and their bottom-up inventory matches more closely with NEI 2011. While not a perfect comparison to BC in 2008, the sign of error in NEI05 relative to the coastal posterior and that study is consistent. An inventory with sector-specific break downs of BC emissions, and additional inversions with more thorough speciated local observations, and higher resolution would all be required to investigate sector-specific anthropogenic pollution.

3.3.4 Error diagnostics

Analysis of posterior emissions uncertainties is useful for understanding the value of the posterior emissions themselves. The diagonal terms of $\mathbf{P}^a$ are the posterior variances, $\boldsymbol{\sigma}_{x_a}$, which are always smaller than prior variances. The variance reduction could instead be presented in β space, by utilizing Eq. 3.16. However, ${\sigma_{\beta^k,i}}^2 < {\sigma_{\beta^0,i}}^2$ is not guaranteed when $x_{a,i} > x_{b,i} = 0$, because the posterior relative emission uncertainty depends on $x_{a,i}$. For this work, the reductions in variance are presented in CV space. The low-rank estimate of $\mathbf{P}^a$ is only valid for linear perturbations away from $\boldsymbol{x}_a$. The final outer loop estimate of $\mathbf{P}^a$ is the most accurate, since it is linearized around the state preceeding $\boldsymbol{x}_a$. A quantitative measure of error reduction in the k_o^{th} outer loop in the i^{th} CV is

$$\rho_{i,k_o} = 1 - \left(\frac{\mathbf{P}^a_{i,i}}{\mathbf{B}_{i,i}}\right)_{k_o} \in [0,1). \tag{3.32}$$

Values of ρ_{i,k_o} closer to 1 reflect locations where the observations provide a stronger constraint than the prior. This estimate may not reflect the entire error reduction, since it does not capture potential reductions in previous outer loops. Without propagating updated estimates for $\mathbf{B}$ to subsequent outer loops (e.g., Tshimanga et al., 2008), we also define ρ_{agg}, a qualitative metric that accounts for increases in curvature (decreases in error) in all outer loops:

$$\rho_{i,\text{agg}} = 1 - \prod_{k_o=1}^{k} \left(\frac{\mathbf{P}^a_{i,i}}{\mathbf{B}_{i,i}}\right)_{k_o} \in [0,1). \tag{3.33}$$

$\rho_{i,\text{agg}}$ reveals additional information about observation footprints not shown by $\rho_{i,k_o=6}$. The non-linear nature of the problem means $\rho_{i,\text{agg}}$ is not quantitative.

Both ($\rho_{k_o=6}$) and ρ_{agg} are presented in Figs. 3.13 and 3.12 for the BB and anthropogenic members of $\boldsymbol{x_a}$, respectively. 50 inner loop iterations were taken in the final outer loop to improve ρ estimates. $\rho_{k_o=6}$ is $< 45\%$ across all scenarios, except for QFED_STD BB sources near the IMPROVE sites on 23/24 June. If the inner loop were halted at 10 iterations, the error reduction estimates are reduced by up to $\sim$10% (i.e., 35% instead of 45%) in the darkest grid cells. Further decreasing uncertainty would require observing the same phenomena more thoroughly, either for

longer periods, with greater spatial coverage, or with more instruments. The BB error reduction shown in Figure 3.13 has similar spatial distributions for FINN_STD and QFED_STD scenarios, but differs significantly between the two time periods due to the different spatial coverage of the observations. The reductions in the north on 22 June are more disperse for QFED_STD, which could be caused by the same regridding errors and plumerise differences that influence the posterior emissions. There is also more error reduction in the south for the QFED_STD emissions. In general, the grid-scale uncertainty improvement is confined to sources close to the observations.

The most obvious application of ρ is to evaluate the footprint of a set of measurements. For example, the large relative BB emission increments in EA1-EA3 on 23/24 June indicate that distant observations can have a large impact on the posterior emissions magnitudes. However, $\rho_{i,k_o=6}$ in Fig. 3.13 indicates there is nearly zero uncertainty reduction for those emissions. Also, upon considering the last two columns of Table 3.6, one might conclude that there is a missing weekend (22 June) to weekday (23/24 June) variation in BC emissions within EA8-10. However, Fig. 3.12 shows that the 22 June observations only weakly reduce uncertainty in emissions.

In a more tangible application, ρ can be used to assess existing and and future observing strategies in a similar way to how Yang et al. (2014) used adjoint sensitivity information to plan future meteorological observing sites to improve forecasts of extreme dust events in the Korean peninsula. Fig. 3.12 presents anthropogenic ρ for different combinations of surface and aircraft observations on 23/24 June. The surface observations primarily resolve sources near Fresno, and to a lesser extent near Los Angeles. Since the purpose of the IMPROVE network is to measure background concentrations, it is mostly successful on 23 June in not being influenced by anthropogenic sources of BC from the major cities. If the goal were to measure anthropogenic sources, inflows, or domain-wide concentrations on daily time scales, then ρ would suggest using a different surface network distribution. Such a conclusion does not conflict with the success of using IMPROVE observations to provide top-down constraints on both BB and anthropogenic emissions on monthly time scales (e.g., Mao et al., 2015).

Another piece of information useful for comparing observing configurations and inversion

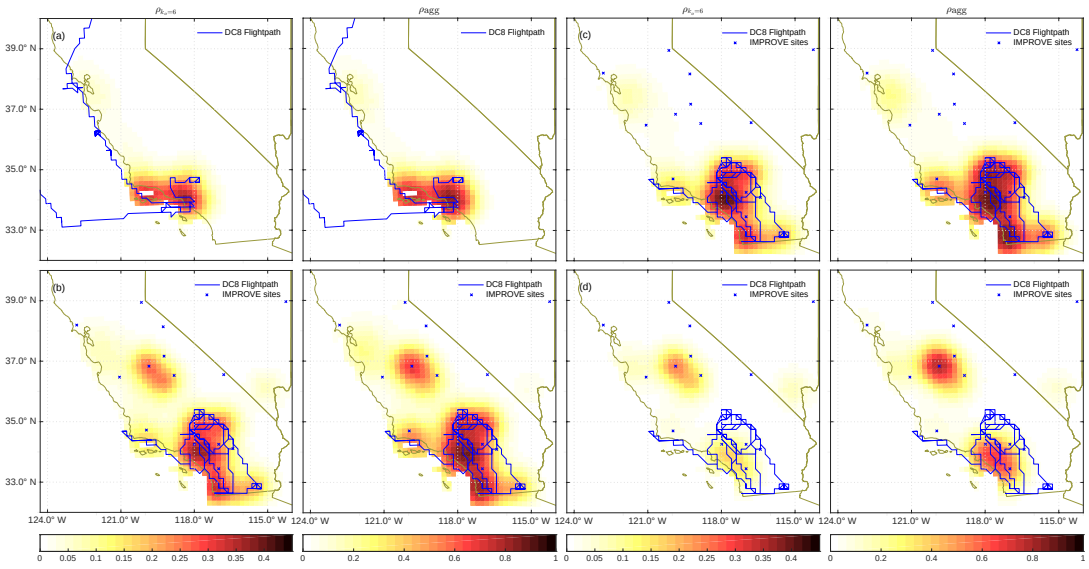

Figure 3.12: Anthropogenic error reduction in the final outer loop ($\rho_{k_o=6}$) and aggregated across all outer loops (ρ_{agg}) for the (a) FINN_STD (22), (b) FINN_STD (23/24), (c) ACFT, and (d) SURF inversion scenarios. The ARCTAS-CARB DC8 flightpath and IMPROVE sites at model grid centers are overlaid.

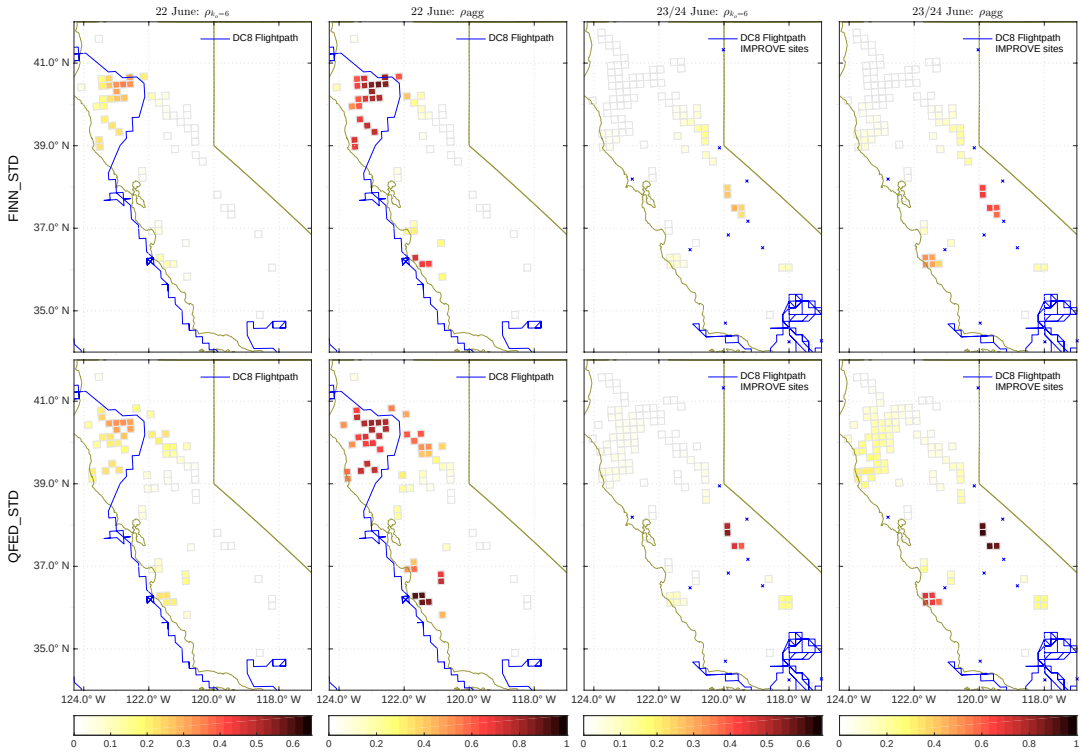

Figure 3.13: BB error reduction in the final outer loop ($\rho_{k_o=6}$) and aggregated across all outer loops (ρ_{agg}) for the two primary BB scenarios on 22 June 00Z-23Z and 23 June, 00Z - 24 June 23Z. The ARCTAS-CARB DC8 flightpath and IMPROVE sites at model grid centers are overlaid.

scenarios is the trace of the resolution matrix, or degrees of freedom for signal, i.e.,

$$\text{DOF} = \text{Tr}\left[\mathbf{I}_n - \mathbf{P}^a\mathbf{B}^{-1}\right], \tag{3.34}$$

which is equal to the number of modes of variability in the emissions that are resolved by the observations (Wahba, 1985; Purser and Huang, 1993; Rodgers, 1996). Substituting the approximation for $\mathbf{P}^a$ from Eq. 3.26,

$$\begin{aligned}\text{DOF} \approx & n - \text{Tr}\left[\left(\mathbf{B} + \mathbf{U}\left(\sum_{k_i=1}^{l}\left(\lambda_{k_i}^{-1} - 1\right)\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^{\top}\right)\mathbf{U}^{\top}\right)\mathbf{B}^{-1}\right] \\ \approx & -\text{Tr}\left[\mathbf{U}\left(\sum_{k_i=1}^{l}\left(\lambda_{k_i}^{-1} - 1\right)\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^{\top}\right)\mathbf{U}^{\top}\mathbf{B}^{-1}\right].\end{aligned} \tag{3.35}$$

Since $\mathbf{U}$ is square, $\mathbf{U}^{\top}\mathbf{B}^{-1}\mathbf{U} = \mathbf{I}_n$, and $\text{Tr}\left[\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^{\top}\right] = \hat{\boldsymbol{\nu}}_{k_i}^{\top}\hat{\boldsymbol{\nu}}_{k_i} = 1$, the expression simplifies as

$$\begin{aligned}\text{DOF} = & -\text{Tr}\left[\left(\sum_{k_i=1}^{l}\left(\lambda_{k_i}^{-1} - 1\right)\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^{\top}\right)\mathbf{U}^{\top}\mathbf{B}^{-1}\mathbf{U}\right] \\ \approx & \sum_{k_i=1}^{l}\left(1 - \lambda_{k_i}^{-1}\right)\text{Tr}\left[\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^{\top}\right] \\ \approx & \sum_{k_i=1}^{l}\left(1 - \lambda_{k_i}^{-1}\right).\end{aligned} \tag{3.36}$$

Therefore, the only information needed to compute DOF are the eigenvalues of $\mathbf{T}_l$. Each inner loop, k_i, has the potential for constraining one additional mode of variability in the emission scaling factors. For all of our inversion scenarios, the leading eigenvalue is on the order of $10^2 - 10^3$, which is equal to the condition number of the full-rank Hessian. As the Lanczos optimization proceeds, each subsequent λ_{k_i} is smaller, asymptotically approaching unity, and each eigenmode provides less information than the one preceding it about scaling factor variability.

Figure 3.14 gives three estimates of DOF at each level of truncation in the final outer loop, that is if higher degrees of eigenvalues were ignored. In that figure, we plot eigenvalue spectra of the FINN_STD and QFED_STD scenarios on 22 June. Similar to ρ, we use a 50 iteration linear optimization to improve the bounds on DOF. The $k_i = l$ estimate of the eigenvalue spectrum at each iteration is represented by a single colored line. Each member of the eigenvalue spectrum,

represented by vertical grid lines in Fig. 3.14, converges toward an upper bound as more iterations are taken. Initial guesses for the least dominant eigenvalues are less than 1 for $k_i \geq 8$ for FINN_STD, but they exceed 1 after an additional iteration, consistent with the properties of the Lanczos sequence. The first DOF in parentheses adheres to the philosophy that only converged eigenvalues should be used to estimate DOF; it excludes $\lambda_{k_i}, \ldots, \lambda_l$ such that λ_{k_i} is more than 5% changed from the previous estimate. The second DOF in parentheses uses all of the current estimates of the eigenvalues available in iteration l. This is still a conservative estimate of DOF, because the true eigenvalues of the full-rank $\mathbf{T}_n$ are always larger than their current numerical estimate. After enough iterations, the numerical growth in DOF is very small, and further computation is not warranted. As the eigenvalue spectra in Figure 3.14 and the cost function reduction in Figure 3.3 show, this is long after the cost function is converged enough for practical purposes. The posterior CVs, which are the primary result from inverse modeling, do not change significantly in the final outer loop. Finally, the best estimates of DOF in red brackets are evaluated at different truncations using the most-converged values of the eigenvalues found in the 22$^{\text{nd}}$ iteration.

Similar to ρ, the quantitative application of DOF is limited to the final outer loop, when $\delta \boldsymbol{x}^n$ is small enough that $(\mathcal{H}_{\delta \boldsymbol{v}})^{-1}|_{x^{n-1}} \approx (\mathcal{H}_{\delta \boldsymbol{v}})^{-1}|_{x^n}$. Absent the need to estimate the posterior Hessian, the outer loop could be ended an iteration earlier. In the inner loop, truncated estimates of $\mathcal{H}^{-1}$ and its eigenvalue spectrum at earlier iterations will provide conservative values for both DOF and ρ. The actual DOF is higher than any value shown in Figs. 3.14 (22 June) and 3.15 (23/24 June). Therefore, the 22 June observations constrain >14 modes of hourly grid-scale variability through 4D-Var in both the FINN_STD and QFED_STD scenarios. Just like for ρ, the optimization constrains additional modes in the earlier outer loop iterations, but that quantification is not straightforward since DOF is defined for linear behavior. If all outer loops were similar, then the total DOF for the entire nonlinear optimization is on the order of 30 to 40.

As shown in Fig. 3.15, the DOF on 23 and 24 June after 50 iterations are 10, 17, and 23 for the SURF, ACFT, and FINN_STD(23/24) scenarios, respectively. The relative magnitudes show that using combined surface and aircraft observations provides an additional value over using

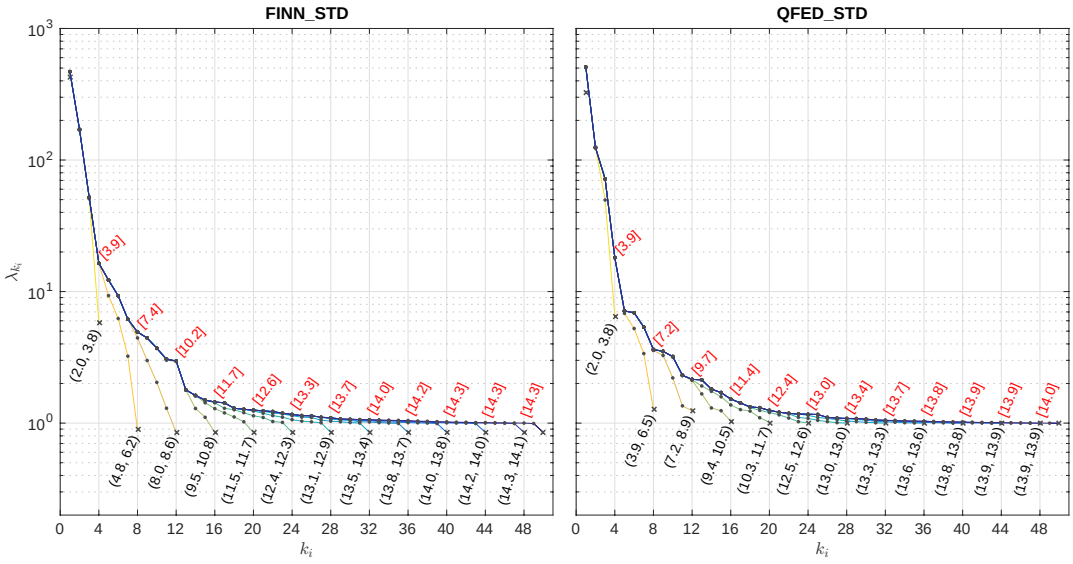

Figure 3.14: Eigenvalue spectra for FINN_STD and QFED_STD in the final outer loop on 22 June. The lines show the estimate of the spectrum $[\lambda_1, \ldots, \lambda_{k_i=l}]$ in every fourth inner loop iteration, l. The black numbers in parentheses are the estimates of DOF that include eigenvalues in the sets (converged to within 5% of the previous estimate, all available). The red numbers in brackets are the truncated estimates of DOF using the most completely converged set of eigenvalues available in the 50$^{\text{th}}$ iteration.

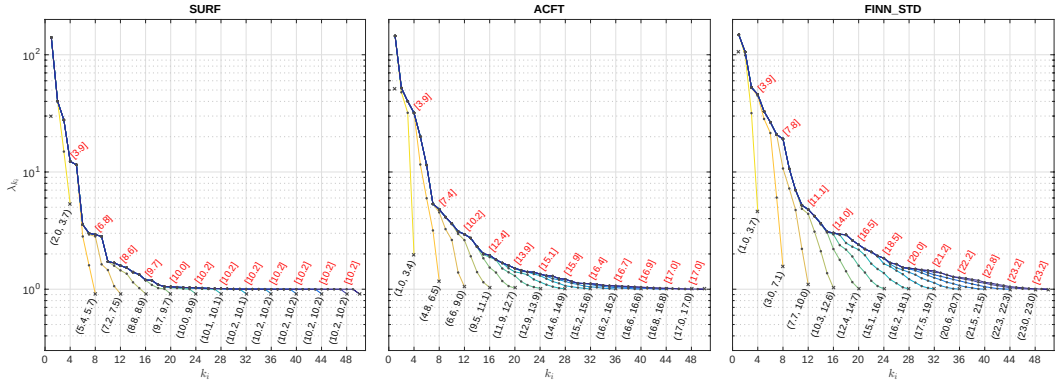

Figure 3.15: Eigenvalue spectra for SURF, ACFT, and SURF+ACFT in the final outer loop on 23 and 24 June. The lines show the estimate of the spectrum $[\lambda_1, \ldots, \lambda_{k_i=l}]$ in every fourth inner loop iteration, l. The black numbers in parentheses are the estimates of DOF that include eigenvalues in the sets (converged to within 5% of the previous estimate, all available). The red numbers in brackets are the truncated estimates of DOF using the most completely converged set of eigenvalues available in the 50$^{\text{th}}$ iteration.

either independently, although the two platforms might have some redundancy. This conclusion is consistent with the maps of BB and anthropogenic ρ in Fig. 3.12, where the footprints of SURF and ACFT have slight overlap near Los Angeles, but are otherwise independent. Additionally, the higher DOF of ACFT is consistent with its more widespread and larger magnitude ρ values. The slower eigenvalue convergence when both observing types are utilized means that additional inner iterations could yield higher estimates for DOF in that case. What is even more clear, and intuitive, is that ρ and DOF estimates require more iterations as the number of constrained CVs increases, which is directly dependent on the number of observations. What might be of particular interest to future measurement planning is that daily average surface data can provide a useful constraint when captured near sources on the same day as the emission event, which is supported by the sparse ρ map for SURF in Fig. 3.12, and the large spike near Fresno.

3.3.5 Cross Validation

As an additional evaluation of the robustness of the emission scaling factors, we apply them in cross validation tests. In two separate evaluations, the 22 June scaling factors are applied to 23-26 June emissions, and the 23/24 scaling factors are applied to 25-26 June emissions. Even before carrying out such a test, the heterogeneous adjoint sensitivity signs and magnitudes for each source sector we found on each day of the campaign (Guerrette and Henze, 2015) are an indication that corrective scaling factors in each day will be unique. In that work, we found that the 24 June observations were most sensitive to Southern California anthropogenic sources on 24 June and to Northern and Southern California coastal sources of both sectors on 23 June. The 26 June observations were most sensitive to Northern California fires, and the adjoint sensitivities were of opposite sign than on 23 and 24 June.

As shown in Fig. 3.4, the cross validated 22 June scaling factors rarely generate improvements to model performance, when compared to 24 and 26 June aircraft observations. On 24 June, some of the high bias predictions are corrected, or even over-compensated, but the low bias prior locations are unaffected. Table 3.3 shows the R^2 and slope of the linear trend lines. The scatter of the fit

for QFED_STD on 24 June and the slope for FINN_STD on 26 June are slightly improved, but all other metrics degrade. The increase in slope for FINN_STD comes as a result of better fit to very large concentrations above the PBL associated with fire sources on multiple previous days. The posterior scaling factors generated from the 23/24 June inversion degrade the forecast of aircraft measurements on 26 June. Since the posterior primarily serves to reduce coastal anthropogenic and BB emissions, it is not surprising that it does not improve a low bias prior two days later.

Table 3.4 includes cross-validated surface measurements on 23 June and 26 June. There is very little change to the modeled surface concentrations as a result of posterior scaling factors in any inversions that only use aircraft observations. Assimilating surface observations on 23 June (Monday) does improve model comparisons to surface observations on 26 June (Thursday). Those small improvements imply that errors are weakly correlated between weekdays. Although it is beyond the information content provided by the observations used in this work, future studies could compare the efficacy of using weak multiday correlation in $\mathbf{B}$ and the hard constraint of 24 h periodic scaling factors used herein.

Given the differing flight tracks on multiple days, the cross validation results demonstrate the need to repeat observations of similar phenomena. Such a strategy could help eliminate non-emission related sources of uncertainty, and further characterize temporal heterogeneity of inventory errors. Aircraft and surface observations do not appear to be useful for cross-validation of each other over the short timescales and limited set of flights considered here. At least for this study period, when they are not collocated, each provides some unique information to the inversion. Cross-validation might be more successful when using measurements collected over a broader range of prior error behaviors or by considering a less complex problem than California statewide BB emissions.

3.4 Conclusions and future work

We have presented the implementation and an application of incremental chemical 4D-Var using an atmospheric chemistry model with online meteorology in WRFDA-Chem. This work

expands on our previous efforts to develop the ADM and TLM in WRFPLUS-Chem (Guerrette and Henze, 2015). This new inversion tool takes advantage of previous developments of meteorological data assimilation in WRFDA (Barker et al., 2005; Huang et al., 2009). That same framework is applied to lognormally distributed emission scaling factors through an exponential transform. We utilize the square root preconditioner for a CVT using horizontal and temporal scaling factor correlations. The Lanczos linear optimization algorithm in the inner loop allows for estimation of posterior error and DOF for objectively evaluating observing systems. Outer loop convergence is improved with a heuristic DGN multiplier, which allows the incremental 4D-Var framework to handle the nonlinearity of the lognormal cost function. While the optimizations herein focus exclusively on emissions, which are known to be important drivers of model uncertainty in BC estimates (e.g., Fu et al., 2012; Zhang et al., 2014a), other factors such as meteorology, plume rise and deposition mechanisms may also affect the model's predictions of BC concentrations.

When applied to the ARCTAS-CARB campaign period, it is not clear which prior emissions perform better. If assessment by initial cost function value alone were meaningful, FINNv1.0 performs best. However, that could be due to FINNv1.0 being biased low combined with the assumption of Gaussian distributed model-observation errors. Positive residuals are weighted higher than negative ones, even when relative errors are equal. There could be some improvement to the posterior emissions by implementing the incremental log-normal cost function framework derived by Fletcher and Jones (2014). If the purpose of the inventory is to provide air quality warnings to the major California cities, then FINNv1.0, FINNv1.5, and QFEDv2.4r8 all have some built-in high bias that will err on the side of caution. Their inability to reproduce high concentrations near sources either points to a deficiency in the inventories, vertical mixing processes, or the temporal observation averaging procedure followed herein, diagnosis of which would require measurements of plume injection heights and widths. The relative magnitudes of grid-scale fire and anthropogenic emissions make it difficult to simultaneously constrain them without additional information. More work should be done to improve both bottom-up and top-down estimates of anthropogenic emissions outside of fire events. We also agree with Mao et al. (2015), who recommended multi-species

inversions (e.g., BC and CO) to discern specific source sectors.

Through the setup and application of the 4D-Var system, we gained valuable knowledge to guide future modeling and measurement efforts. We found two errors in the diurnal distribution of BB emissions and identified a scaling necessary to apply QFED to the western U.S. Additionally, the highly heterogenous posterior scaling factors during ARCTAS-CARB raise questions that the limited BB observations during that time period do not answer. (1) Are BB emission errors always heterogeneous, or only during a transient initiation stage like that observed in June 2008? If heterogeneity is consistent outside initiation events, then inversions should apply weaker inter-day correlation than the hard constraint used herein or have independent scaling factors for each day. (2) Are the temporally bimodal posterior emissions realistic, or are they an artifact of the correlation timescale used? (3) Are the BB plume heights reasonable, and should they follow a diurnal pattern? The current 1D plume rise mechanism in WRF-Chem depends strongly on specified burned areas, which are diurnally invariant and highly uncertain (e.g., Boschetti et al., 2004). The last two questions indicate there is value in continuous night (between 20:00 and 06:00 LT) and day measurements of the same fire region. Since models poorly predict shallow boundary layers, the use of night time observations in 4D-Var would require characterization and subsequent model tuning of those vertical mixing processes. Furthermore, if it is accepted that high-resolution models are required to accurately predict degraded air quality events, then high spatial and/or temporal resolution concentration measurements from research campaigns or geostationary satellites are necessary to provide the sufficient constraints on inventory errors. The error reduction estimation method provided herein will be useful for planning these future missions.

Future applications of the WRFDA-Chem system developed here may consider improvements such as the following. One possible way to reduce model uncertainty would be to extend the multi-incremental 4D-Var available in WRFDA (Zhang et al., 2014b) to the new scaling factor CVs. Multi-incremental chemical 4D-Var would use a high-resolution model forecast to generate trajectory checkpoint files (see Guerrette and Henze (2015)), and could take advantage of improvements to chemical transport at higher resolution realized by using online meteorology demonstrated by

Grell et al. (2004) and Grell and Baklanov (2011). In addition, FDDA nudging has been shown to improve wind fields and was used successfully in an LPDM emission inversion (Lauvaux et al., 2016). Even after exhausting methods to improve the posterior, the error contributions from hard-coded descriptions of meteorology can be bounded using ensemble and sensitivity tests (e.g., Angevine et al., 2014; Lauvaux et al., 2016).

Chapter 4

A New Randomized Incremental Optimal Technique (RIOT) for Four Dimensional Variational Data Assimilation

4.1 Introduction

Incremental four dimensional variational (4D-Var) data assimilation (DA) is used in operational numerical weather prediction (NWP) around the world (e.g., European Center for Medium-Range Forecasts, UK Met Office, Meteo France, Panasonic Weather Solutions, U.S. Naval Research Laboratory). In Chapter 3, we applied this method, with some modification, to the inversion of chemical emissions. A defining drawback of 4D-Var relative to its ensemble-based counterparts in atmospheric DA has been the necessity to integrate a forward model (FWM) and its adjoint (ADM) repeatedly in a sequential algorithm. Each serial calculation is usually parallelized across many processors; however, this type of parallelization is limited by communication bottlenecks and hardware constraints. In contrast, ensemble approaches are embarrassingly parallel, as their numerous forward model evaluations can be evaluated simultaneously and independently.

There have been several efforts to reduce the wall-time in 4D-Var either through reduction of floating point operations or through parallelization. The wall-time for each sequential iteration can be reduced through multi-incremental 4D-Var, which runs the expensive tangent linear (TLM) and AD integrations on a lower resolution grid than the full model. Multi-incremental 4D-Var is used in operational NWP, and produces comparable results to optimizations that use full-scale grids (e.g., Zhang et al., 2014b). Instead of reducing the time of a single iteration, advanced preconditioning methods are aimed at reducing the number iterations required to converge (e.g., Desroziers and

Berre, 2012; Gürol et al., 2014). Trémolet (2006) introduced weak-constraint 4D-Var, in which a temporally segmented analysis window accounts for model uncertainty at the segment boundaries. This method lends itself to parallel integration in time, which was recently realized through a saddle cost function formulation (Fisher et al., 2011). The saddle formulation also requires new preconditioning methods to enable convergence and wall-time reductions in practical applications (Fisher et al., 2011, 2016). Similar parallel-in-time integrations have been implemented in strong-constraint 4D-Var by using an augmented Lagrangian version of the cost function (Rao and Sandu, 2016). Brown et al. (2016) enabled parallelism across spatial modes through multilevel 4D-Var where Hessian modes at discrete spatial scales are computed simultaneously, but still through iteration.

Temporal parallelism holds promise for enabling longer analysis windows that account for more observations simultaneously, but also poses challenges in implementation. It would be additionally beneficial to have an ensemble-like parallelism in adjoint integration where the computational scalability is dependent on the number of modes of variability constrained by the optimization. Hypothetically, such parallelism is complementary to any of the other forms described above if there are enough available processors. In this chapter we apply the recently developed Randomized Incremental Optimal Technique (RIOT) (Bousserez and Henze (2016); hereafter BH16) to exploit the computational advantages of ensemble calculations while maintaining the merits of 4D-Var.

4.1.1 Background

Here we present some mathematical preliminaries on optimization that are important to understanding the contribution of this work. In Appendix A we demonstrated the equivalence between the control variable steps from incremental 4D-Var and those from the Gauss-Newton (GN) method. In GN, the cost function is linearized around a state, $\boldsymbol{v}_0 \in \mathbb{R}^n$, i.e.,

$$J\left(\boldsymbol{v}_0 + \delta\boldsymbol{v}\right) = \delta\boldsymbol{v}^\top \mathbf{A}\delta\boldsymbol{v} + \delta\boldsymbol{v}^\top \boldsymbol{b} + J\left(\boldsymbol{v}_0\right). \tag{4.1}$$

Its gradient,

$$\nabla J = \mathbf{A}\delta\boldsymbol{v} + \boldsymbol{b}, \tag{4.2}$$

is set equal to zero to solve for an optimal increment,

$$\delta\boldsymbol{v}^* = -\mathbf{A}^{-1}\boldsymbol{b}, \tag{4.3}$$

in terms of the gradient and Hessian (second derivative) evaluated at $\boldsymbol{v}_0$, i.e., $\boldsymbol{b} = \frac{\partial J}{\partial \boldsymbol{v}}|_{\boldsymbol{v}_0} \in \mathbb{R}^n$ and $\mathbf{A} = \frac{\partial^2 J}{\partial \boldsymbol{v}^2}|_{\boldsymbol{v}_0} \in \mathbb{R}^{n\times n}$, respectively. There are several methods to account for nonlinearities that cause the increment to be non-optimal (e.g., Levenberg-Marquardt, Quasi-Newton methods, trust region), but herein we apply a damping multiplier as

$$\delta\boldsymbol{v}^* = -\eta\mathbf{A}^{-1}\boldsymbol{b}. \tag{4.4}$$

$\eta \in \mathbb{R} \,|\, 0 < \eta \leq 1$ is found by a line search, or 1-D optimization on values of J, once $\mathbf{A}^{-1}$ and $\boldsymbol{b}$ are known. After updating $\boldsymbol{v}^{k+1} = \boldsymbol{v}^k + \delta\boldsymbol{v}^k$, the system of equations is relinearized around the new state and the iterative process continues until a convergence criteria is reached or until a specified number of iterations, k_f, are complete.

When GN is applied to large-scale atmospheric problems, there are $n \gtrsim 10^5 - 10^6$ control variables and $m \gtrsim 10^2 - 10^5$ observations to constrain them, such that $\text{rank}\,(\mathbf{A}) \leq \min(m, n)$. $\mathbf{A}$ is not represented explicitly, and finding its inverse is the crux of the problem. To do so, $\mathbf{A}$ is approximated in terms of a low-rank eigendecomposition. Incremental 4D-Var uses symmetric Krylov methods (i.e., conjugate gradient (CG), Lanczos recurrence), which require a second level of iteration (inner loop) to perform the decomposition. Krylov methods require multiplying $\mathbf{A}$ to increasing degrees by $\boldsymbol{b}$; this is a sequential procedure comprised of l matrix vector multiplications, where l is the desired rank of the approximation. As the rank of approximation, l, increases, the error associated with representing $\mathbf{A}$ using only the leading eigenvalues may become smaller than that associated with the GN linearization. Therefore, there is an inner iteration beyond which there is no perceived improvement to the posterior state, $\boldsymbol{v}^{k+1}$, even as the approximation of $\mathbf{A}$ improves.

In machine learning, where the functional relationship that maps the control variables to each observation in a training set is separable from the others, i.e.,

$$J(\boldsymbol{v}) = \frac{1}{m}\sum_{i=1}^{m} y_i(\boldsymbol{v}),\ i = \{1, 2, \ldots, m\}, \tag{4.5}$$

a recent solution is to subsample the Hessian by only using a subset of m (e.g., Byrd et al., 2011; Roosta-Khorasani and Mahoney, 2016). Following such an approach in atmospheric problems would be equivalent to using only a few observations in each iteration, which would likely lower the rank of $\mathbf{A}$ and the number of Krylov iterations required. However, the 4D-Var Hessian is not separable by observations, which precludes using this type of parallelism.

Recently, Woolfe et al. (2008) proposed and demonstrated a randomized singular value decomposition (RSVD) method for approximating a high-dimensional matrix such as $\mathbf{A}$. RSVD is a block algorithm that begins by multiplying $\mathbf{A}$ by a matrix containing a set of vectors drawn from a Gaussian distribution. RSVD is useful for matrix-free decompositions, where the matrix of interest is represented by a set of equations, but its members are not explicitly represented. In general, block algorithms are easily parallelized because each matrix-vector multiplication is independent. Clarkson and Woodruff (2009) further decreased the wall-time of RSVD by introducing single-pass methods that circumvents a second multiplication by $\mathbf{A}$. Martinsson et al. (2011) showed the importance of oversampling, or more random vectors $(l + p)$ than the desired rank of the approximation, l. This is akin to Krylov algorithms that require additional iterations to converge on earlier eigenmodes that are not fully constrained. Halko et al. (2011) (hereafter HMT11) reviewed and synthesized many algorithmic variants of RSVD and gave error approximations for each one. Those authors' work is the basis for the methods applied herein.

There are several relevant studies that have applied RSVD or similar approaches, but which are slightly different from the work herein. Liu et al. (2015) provided an alternative block algorithm for finding the low-rank approximation of a non-square matrix. Those authors use a GN optimization within that matrix approximation, but they do not apply the resulting low-rank matrix in a GN procedure as we are intending to do. Hsieh and Olsen (2014) similarly use RSVD

within their own Newton-based approach to matrix active subspace determination, i.e., low-rank approximation, but do not conduct GN using their result. Kitanidis and Lee (2014) apply RSVD to approximate the Hessian in a geostatistical GN algorithm that uses an adjoint-free ensemble of FWM integrations. They stipulate that their algorithm is limited to applications where "the measurements have limited information content," which would be true when m is small or the observations are highly correlated.

The underlying theory and algorithm for our work is given by BH16. BH16 applied adjoint-based RSVD to approximate the incremental 4D-Var Hessian in the first GN iteration of an atmospheric surface flux inversion problem and compare the analysis increment to that from forty iterations of BFGS. They also characterized the convergence of standard and optimal formulations for the approximate Hessian. After these initial successes, they propose the Randomized Incremental Optimal Technique (RIOT) to solve nonlinear optimization problems across multiple GN iterations.

4.1.2 Summary of this work

Herein we implement RIOT, and apply it to a regional chemical source inversion problem in the Weather Research and Forecasting Data Assimilation with chemistry (WRFDA-Chem) tool (Chapter 3). We use a lognormal distribution for surface fluxes (described in Sec. 3.2.3.5), which means it is important to account for nonlinearities through the multiple outer loop framework of GN. We demonstrate that the parallelized Hessian approximation has computational and accuracy benefits over traditional methods that will improve the competitiveness of adjoint-based DA. In Sec. 4.2, we summarize how a Lanczos-GN minimization and RIOT approximate the Hessian of the 4D-Var cost function, and how these are used to calculate analysis increments and the posterior covariance. We also describe two objective metrics for evaluating matrix approximations and their implied analysis increments when an exact evaluation of the Hessian is available. In Sec. 4.3, we compare eigenmodes and analysis increments in the first outer iteration of incremental 4D-Var using (1) a truncated SVD of the exact Hessian, (2) the Lanczos recurrence, and (3) several variations of

RIOT. We also evaluate converged states and posterior variance from a Lanczos-GN optimization and RIOT after several outer iterations. In Sec. 4.4, we summarize the utility of RIOT and give insight into it's applicability in NWP.

4.2 Methods

4.2.1 Incremental 4D-Var

As we described in Sec. 3.2.3.1, and following the same notation, we apply GN to the incremental 4D-Var using a square-root preconditioner, i.e.,

$$\begin{aligned} J\left(\delta\boldsymbol{v}^k\right) =& \frac{1}{2}\left(\delta\boldsymbol{v}^k - \boldsymbol{d}^{b,k-1}\right)^\top \left(\delta\boldsymbol{v}^k - \boldsymbol{d}^{b,k-1}\right) \\ &+ \frac{1}{2}\left(\mathbf{G}^{k-1}\mathbf{U}\delta\boldsymbol{v}^k - \boldsymbol{d}^{o,k-1}\right)^\top \mathbf{R}^{-1}\left(\mathbf{G}^{k-1}\mathbf{U}\delta\boldsymbol{v}^k - \boldsymbol{d}^{o,k-1}\right). \end{aligned} \tag{4.6}$$

$\delta\boldsymbol{v} \in \mathbb{R}^n$ is a preconditioned increment; $\mathbf{U}$ is the square-root preconditioner ($\delta\boldsymbol{x} = \mathbf{U}\delta\boldsymbol{v}$, $\mathbf{B} = \mathbf{U}\mathbf{U}^\top$); $\delta\boldsymbol{x}$ is an increment in the control variable (CV) space, $\boldsymbol{x}$; $\mathbf{G}$ is the linearized model operator, or Jacobian of the system of equations; $\mathbf{R}$ is the model-observation error covariance matrix; and $\mathbf{B}$ is the prior error covariance of the CV, $\boldsymbol{x}$. The observation and background innovations are

$$\boldsymbol{d}^{o,k-1} = \boldsymbol{y}^o - G\left(\boldsymbol{x}^{k-1}\right), \tag{4.7}$$

and

$$\boldsymbol{d}^{b,k-1} = \sum_{k_o=1}^{k-1} \delta\boldsymbol{v}^{k_o}, \tag{4.8}$$

respectively, where $\boldsymbol{y}^o \in \mathbb{R}^m$ is the set of observations. $\mathbf{G}$ and its transpose have superscripts of $k-1$ to indicate linearization about the CV from the previous outer iteration, which we will drop from this point on for readability. The analysis increment calculated in the k^{th} outer iteration of incremental 4D-Var is

$$\begin{aligned} \delta\boldsymbol{v}^k &= \left(\mathbf{I}_n + \mathbf{U}^\top\mathbf{G}^\top\mathbf{R}^{-1}\mathbf{G}\mathbf{U}\right)^{-1}\left(\boldsymbol{d}^{b,k-1} + \mathbf{U}^\top\mathbf{G}^\top\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right) \\ &= -\left[\mathcal{H}_{\delta\boldsymbol{v}}\right]^{-1}\nabla_{\delta\boldsymbol{v}}J|_{\delta\boldsymbol{v}^k=\mathbf{0}}, \end{aligned} \tag{4.9}$$

in preconditioned CV space. The second term, a vector, in the product of Eq. 4.9 is the gradient of the cost function evaluated at the most recent set of CVs, which requires an adjoint simulation.

We showed in Appendix B that the first term in the product of Eq. 4.9, the inverse of the preconditioned Hessian, can be represented as a low-rank *update* (LRU) to the prior preconditioned covariance (identity matrix, $\mathbf{I}_n$) in terms of the leading l eigenmodes (eigenvalues, $\mathbf{\Lambda}_l = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_l)$; eigenvectors,$\mathbf{W}_l = [\hat{\boldsymbol{\nu}}_1, \hat{\boldsymbol{\nu}}_2, \ldots, \hat{\boldsymbol{\nu}}_l]$) of the preconditioned Hessian, i.e.,

$$\mathcal{H}_{\delta \boldsymbol{v}} \approx \mathbf{I} + \sum_{k_i=1}^{l} \left(\lambda_{k_i} - 1\right) \hat{\boldsymbol{\nu}}_{k_i} \hat{\boldsymbol{\nu}}_{k_i}^{\top},$$

and its inverse (see Eq. B.12)

$$[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1} \approx \mathbf{I} + \sum_{k_i=1}^{l} \left(\lambda_{k_i}^{-1} - 1\right) \hat{\boldsymbol{\nu}}_{k_i} \hat{\boldsymbol{\nu}}_{k_i}^{\top}, \tag{4.10}$$

or equivalently,

$$\text{LRU: } [\mathcal{H}_{\delta \boldsymbol{v}}]^{-1} \approx \mathbf{I} - \sum_{k_i=1}^{l} \left(\frac{\lambda_{k_i} - 1}{\lambda_{k_i}}\right) \hat{\boldsymbol{\nu}}_{k_i} \hat{\boldsymbol{\nu}}_{k_i}^{\top} \equiv \mathbf{P}_{\boldsymbol{v},\text{U}}^{a} \approx \mathbf{P}_{\boldsymbol{v}}^{a}, \tag{4.11}$$

where $l << n$, and $\text{rank}\left(\mathcal{H}_{\delta \boldsymbol{v}}\right) \leq \min\left(m, n\right)$. k_i is the iterator across spectral modes. For sequential methods, this refers to the iterator, while for parallel methods, this refers to the counter of ensemble members. As we discussed in Sec. 3.2.3.4, $[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1}$ is equivalent to the preconditioned posterior covariance, $\mathbf{P}_{\boldsymbol{v}}^{a}$, near a stationary point in the optimization. BH16 state that without any modification, CG and the Lanczos recurrence do not use an LRU, but instead intrinsically apply a low-rank *approximation* (LRA) for the inverse Hessian, i.e.,

$$\text{LRA: } [\mathcal{H}_{\delta \boldsymbol{v}}]^{-1} \approx \sum_{k_i=1}^{l} \frac{1}{\lambda_{k_i}} \hat{\boldsymbol{\nu}}_{k_i} \hat{\boldsymbol{\nu}}_{k_i}^{\top} \equiv \mathbf{P}_{\boldsymbol{v},\text{A}}^{a} \approx \mathbf{P}_{\boldsymbol{v}}^{a}. \tag{4.12}$$

In principle, the LRU can be used within the Lanczos recurrence if the standard increment algorithm (e.g., Golub and Van Loan, 1996) is replaced with an explicit formulation that uses $\mathbf{\Lambda}$ and $\mathbf{V}$. We will show in Sec. 4.3.1 that when the Lanczos recurrence is used to calculate an analysis increment, there is no distinction between the LRU and LRA. For further explanation of this issue, see Sec. 4.2.2.3.

Once $\delta \boldsymbol{v}$ is known, it is easily converted to CV space by applying the preconditioner, i.e., $\delta \boldsymbol{x}^k = \mathbf{U} \boldsymbol{v}^k$. Following from Eq. 3.26, and as we confirmed in Appendix C, the posterior covariance

is converted to CV space through

$$\mathbf{P}^a = \mathbf{U}\mathbf{P}_{\boldsymbol{v}}^a\mathbf{U}^\top. \tag{4.13}$$

The posterior variance is used in atmospheric chemistry modeling to decide whether observations give statistically significant information about emission sources or initial states at varying temporal and spatial scales (see Chapter 3 for an example). The practical differences for analysis increments and posterior variance between the LRU and LRA are explored in Sec. 4.3.

Spantini et al. (2015) showed optimality results for analysis increments calculated using either the LRU or the LRA. BH16 repeated those derivations and presented new optimality results for the posterior covariance. We reiterate the analysis increment optimality conditions here, because we use them later in the paper. Those results depends on two different matrix classes, as described by BH16:

$$\mathcal{A}_l \equiv \{\mathbf{M} \in \mathcal{M}_n | \operatorname{rank}(\mathbf{M}) \leq l\}$$
$$\hat{\mathcal{A}}_l \equiv \left\{\mathbf{M} \in \mathcal{M}_n | \mathbf{M} = \mathbf{B} - \mathbf{Q}\mathbf{Q}^\top \geq 0, \operatorname{rank}(\mathbf{Q}) \leq l\right\}$$

$\mathcal{A}_l$ is the class of $n \times n$ matrices which are referred to as LRA's. $\hat{\mathcal{A}}_l$ is the class of negative semidefinite updates to the prior error covariance matrix $\mathbf{B}$, where $\mathbf{Q}$ is a general low-rank matrix. Approximations belonging to this class are referred to as LRU's.

BH16 showed that the Bayes risk of the analysis increment calculated with the LRU and LRA are

$$\mathbb{E}\|\boldsymbol{x}_{\mathrm{LRU}} - \boldsymbol{x}^a\|^2_{(\mathbf{P}^a)^{-1}} = \min_{\widetilde{\mathbf{P}} \in \hat{\mathcal{A}}_l} \mathbb{E}\| \left(\widetilde{\mathbf{P}} - \mathbf{P}^a\right) \mathbf{G}^\top \mathbf{R}^{-1} \boldsymbol{d}^k\|^2_{(\mathbf{P}^a)^{-1}} = \sum_{i>l} (\lambda_i - 1)^3, \tag{4.14}$$

and

$$\mathbb{E}\|\boldsymbol{x}_{\mathrm{LRA}} - \boldsymbol{x}^a\|^2_{(\mathbf{P}^a)^{-1}} = \min_{\widetilde{\mathbf{K}} \in \mathcal{A}_l} \mathbb{E}\| \left(\widetilde{\mathbf{K}} - \mathbf{K}\right) \boldsymbol{d}^k\|^2_{(\mathbf{P}^a)^{-1}} = \sum_{i>l} (\lambda_i - 1), \tag{4.15}$$

respectively. $\mathbf{P}^a$ is the true posterior covariance, $\boldsymbol{x}^a$ is posterior mean that would be found by using the exact Hessian, $\|\cdot\|_{\mathbf{M}}$ is the weighted Euclidian norm with respect to the matrix $\mathbf{M}$, and $\mathbb{E}(\cdot)$

is the expectation operator. $\boldsymbol{d}^k$ is a combined observation and background innovation evaluated after k outer iterations, and is in observation space. In Eq. 4.15, $\mathbf{K}$ is the Kalman gain matrix, $\mathbf{K} = \mathbf{B}\mathbf{G}^\top \left(\mathbf{G}\mathbf{B}\mathbf{G}^\top + \mathbf{R}\right)^{-1}$. As can be seen from Eqs. 4.14 and 4.15, there is an intersection at $\lambda_{l+1} = 2$ where the LRU becomes more optimal than the LRA for the analysis increment.

It should be noted that BH16 express the Bayes risk, increments, and posterior covariance of the LRU and LRA in terms of the eigenmodes of the preconditioned observation Hessian,

$$\mathcal{H}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-1} \mathbf{G}\mathbf{U}, \tag{4.16}$$

which only differ from those of the full preconditioned Hessian by $\mathbf{\Lambda}_{\mathcal{H}_{\delta \boldsymbol{v}}} = \mathbf{I}_n + \mathbf{\Lambda}_{\mathcal{H}_{\delta \boldsymbol{v},o}}$, where $\mathbf{\Lambda}$ are the respective diagonal matrices of eigenvalues of either matrix. The elements of $\mathbf{\Lambda}_{\mathcal{H}_{\delta \boldsymbol{v}}}$ are all ≥ 1, whereas the elements of $\mathbf{\Lambda}_{\mathcal{H}_{\delta \boldsymbol{v},o}}$ are all ≥ 0. We will be converting between these two Hessians throughout Sec. 4.2.2.2.

BH16 evaluated RSVD approximations of the eigendecomposition of $\mathcal{H}_{\delta \boldsymbol{v}}$ for a small problem ($n = 300$), for which a direct SVD computation was possible through explicit construction of the Hessian using finite-differences of adjoint sensitivities. Those authors also applied RSVD and BFGS to a larger problem ($n = 18271$) to demonstrate the wall-time benefits of the former. Here we use multiple methods to determine the eigendecomposition of $\mathcal{H}_{\delta \boldsymbol{v}}$ for a relatively large problem with few observations ($n \approx 3 \times 10^5$, $m = 241$): the Lanczos recurrence, several forms of RSVD, and a hybrid Krylov-RSVD variant. We assume that the desired rank of the Hessian approximation, l, and the number of outer loops, k_f, are specified. Here we give the details of each of these methods to illustrate their theoretical differences.

4.2.2 Hessian Approximations

4.2.2.1 Lanczos recurrence

Desroziers and Berre (2012) give a thorough description of how the Lanczos algorithm is applied to incremental 4D-Var. We give a summary here to illustrate its sequential nature. The Lanczos recurrence determines a Krylov subspace of a Hermitian target matrix ($\mathbf{A} = \mathbf{A}^* \equiv \mathcal{H}_{\delta \boldsymbol{v}}$)

by sequentially multiplying it by a test vector ($\boldsymbol{b} \equiv \nabla_{\delta \boldsymbol{v}} J|_{\delta \boldsymbol{v}^k = \boldsymbol{0}}$) in increasing powers to determine the range of the target matrix, i.e.,

$$\mathcal{K}_l\left(\mathbf{A}, \boldsymbol{b}\right) = \text{span}\left\{\boldsymbol{b}, \mathbf{A}\boldsymbol{b}, \mathbf{A}^2\boldsymbol{b}, \ldots, \mathbf{A}^{l-1}\boldsymbol{b}\right\}. \tag{4.17}$$

We focus on symmetric matrices here because of the inherent symmetry of $\mathbf{A} = \mathcal{H}_{\delta \boldsymbol{v}}$, but the Krylov subspace of non-symmetric matrices can be found by the Arnoldi algorithm. From here, we assume that $\mathbf{A}$ is a real matrix, which allows us to use the transpose instead of the Hermitian operator. The Lanczos vectors discussed in Sec. 3.2.3.4 and Appendix B form an orthonormal basis for the Krylov subspace. Therefore, $\mathbf{Q}_l = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_l] \in \mathbb{R}^{n \times l}$ forms a basis set for $\mathcal{H}_{\delta \boldsymbol{v}}$. It is well-known that the Lanczos algorithm produces orthogonal $\boldsymbol{q}_{k_i}$'s in exact arithmetic, but not in finite precision. Modified Gram-Schmidt (MGS) is used for re-orthonormalization.

In Appendix B, we describe how the Lanczos recurrence builds the LRA of $\mathcal{H}_{\delta \boldsymbol{v}}$ as

$$\mathcal{H}_{\delta \boldsymbol{v}} \approx \mathbf{Q}_l \mathbf{W}_l \mathbf{\Lambda}_l \mathbf{W}_l^\top \mathbf{Q}_l^\top. \tag{4.18}$$

$\mathbf{W}_l = [\hat{\boldsymbol{w}}_1, \hat{\boldsymbol{w}}_2, \ldots, \hat{\boldsymbol{w}}_l]$ and $\mathbf{\Lambda}_l = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_l)$ comprise the spectral decomposition of a symmetric tridiagonal matrix, $\mathbf{K}_{\text{st},l} \in \mathbb{R}^{l \times l}$, which is an inherent component of Lanczos. Each eigenvector of $\mathcal{H}_{\delta \boldsymbol{v}}$ is $\hat{\boldsymbol{\nu}}_{k_i} = \mathbf{Q}_l \hat{\boldsymbol{w}}_{k_i}$. The diagonal elements of $\mathbf{\Lambda}_l$ are called the Ritz eigenvalues, and have been shown to approximate the extremal eigenvalues of $\mathbf{A}$. Such approximation should improve as $l \to \text{rank}\left(\mathbf{A}\right)$, where in our case $\text{rank}(\mathbf{A}) = \text{rank}(\mathcal{H}_{\delta \boldsymbol{v}}) = n$. That is to say, l iterations of the Lanczos recurrence produces a rank-l approximation of $\mathcal{H}_{\delta \boldsymbol{v}}$, but one that is not identical to the rank-l approximation that would result from an eigendecomposition of the explicitly formed Hessian. The Lanczos recurrence requires p extra iterations, for a total of $l + p$, to match the best possible rank-l approximation of $\mathbf{A}$; p is the oversampling parameter.

4.2.2.2 Randomized SVD

Whereas Krylov subspace methods apply the target matrix in increasing degrees to form the rank-l basis set, $\mathbf{Q}_l$, RSVD (Woolfe et al., 2008) multiplies the target matrix by a set of l randomly

drawn column vectors, $\mathbf{\Omega} = [\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_l]$ to form

$$\mathbf{Y} = \hat{\mathbf{A}}\mathbf{\Omega}. \tag{4.19}$$

When applied to incremental 4D-Var, $\hat{\mathbf{A}}$ is either $m \times n$, $n \times n$, or $n \times m$. In a method we refer to as RIOT-56, $\hat{\mathbf{A}}$ is square, and we set the observation Hessian to $\mathcal{H}_{\delta \boldsymbol{v},o} = \mathbf{A} = \hat{\mathbf{A}}$. In another approach, RIOT-51, $\hat{\mathbf{A}}$ is non-square, and we will use $\mathcal{H}_{\delta \boldsymbol{v},o} = \mathbf{A} = \hat{\mathbf{A}}^\top \hat{\mathbf{A}}$. Herein, the members of $\mathbf{\Omega}$ are independent and identically distributed standard Gaussian variables with mean 0 and standard deviation 1, but HMT11 note that other random matrices achieve similar results. All of the matrix vector products, $\boldsymbol{y}_{k_i} = \mathbf{A}\boldsymbol{\omega}_{k_i}$, $k_i \in (1, 2, \ldots, l)$, are computed in parallel. Although we use the same subscript as we did for the Lanczos recurrence, now k_i identifies a single ensemble member, not an iteration. In this step, RSVD takes the bulk of the computational load from the Lanczos recurrence and spreads it across a number of processes equal to the number of modes of variability in $\mathbf{A}$ that the user would like to approximate. Once agaion, l corresponds to the rank of the eventual matrix approximation, but that approximation will not be identical to the rank-l decomposition that would result from an SVD of an explicit form of $\mathbf{A}$. The solution is to increase l by an oversampling parameter, p, except that now the additional computations are conducted in parallel. From this point, we will assume that $N_{\text{app}} = l + p$ random vectors are used, where N_{app} is the number of approximated modes and l is the effective rank that those modes represent.

Once $\mathbf{Y}$ is found, we proceed with the randomized range finder, Algorithm 4.1 of HMT11. Alternatively, HMT11 Algorithm 4.4 could be used, which we discuss later in this section. The basis of $\mathbf{Y}$, denoted $\mathbf{Q}$, is found through some QR decomposition method (e.g., Gram-Schmidt orthogonalization, Householder matrices, or by using the left singular vectors of $\mathbf{Y}$). Following the recommendation of HMT11, we carry out all QR decompositions using a twice repeated version of MGS, as proposed by Björck (1994). The range of $\mathbf{Q}$ is an l-dimensional subspace that approximates the range of $\hat{\mathbf{A}}$. Therefore, one has the following bounded spectral norm:

$$\| \left[\mathbf{I}_n - \mathbf{Q}\mathbf{Q}^\top \right] \hat{\mathbf{A}} \|_2 \leq \epsilon. \tag{4.20}$$

HMT11 gives expected upper limits for ϵ achievable by RSVD for general matrices. As mentioned previously, RSVD requires N_{app} realizations to achieve a true rank-l approximation of $\hat{\mathbf{A}}$. HMT11 state that a oversampling of $p = 5$ or $p = 10$ should be sufficient, but also that in some rare cases $p = l$ may be necessary.

BH16 propose three different approaches to applying RSVD to the cost function Hessian, each of which (1) defines a unique matrix $\hat{\mathbf{A}}$ of specific dimension, and (2) utilizes either Algorithms 5.1 or 5.6 from HMT11 to determine a low-rank representation of $\hat{\mathbf{A}} = \mathbf{Q}\mathbf{K}\mathbf{Q}^\top$. That decomposition is analogous to the Lanczos decomposition except that $\mathbf{K}$ is no longer tridiagonal. BH16 give a complete description of the RIOT algorithm and how it would be used in operational forecasting. We repeat some derivations here to clarify specific points about the different variations of RIOT. The algorithm descriptions for RIOT using HMT11 Algorithm 5.1 (RIOT-51) or HMT11 Algorithm 5.6 (RIOT-56) are repeated in Appendix D. In all three approaches, we apply RSVD to the observation Hessian, $\mathcal{H}_{\delta \boldsymbol{v},o}$, and then transform to the full Hessian, $\mathcal{H}_{\delta \boldsymbol{v}}$. The three approaches are as follows:

(1) **RIOT-56**: Let $\hat{\mathbf{A}} \equiv \mathcal{H}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-1} \mathbf{G} \mathbf{U} \in \mathbb{R}^{n \times n}$

$\mathbf{Y} \in \mathbb{R}^{n \times N_{\text{app}}}$ is found by multiplying $\hat{\mathbf{A}}$ by random vectors, $\boldsymbol{\omega}_i \in \mathbb{R}^n$, in preconditioned CV space. Once $\mathbf{Q} \in \mathbb{R}^{n \times l}$ is known, it is used in the single-pass Algorithm 5.6 of HMT11, and as proposed by Clarkson and Woodruff (2009), which applies to a symmetric $\hat{\mathbf{A}}$. In short, this algorithm takes utilizes the error bound in Eq. 4.20 to reach the expression

$$\mathbf{K}\mathbf{Q}^\top \boldsymbol{\Omega} \approx \mathbf{Q}^\top \mathbf{Y}. \tag{4.21}$$

A least-squares solver can be used to find a symmetric $\mathbf{K} \in \mathbb{R}^{N_{\text{app}} \times N_{\text{app}}}$. In the case when $\mathbf{U}$ and $\mathbf{H}$ are perfectly modeled as symmetric, the spectral decomposition, $\mathbf{K} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}$, yields the approximate decomposition of $\hat{\mathbf{A}}$ as

$$\hat{\mathbf{A}} \approx \mathbf{Q}\mathbf{K}\mathbf{Q}^\top = \mathbf{Q}\mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^\top\mathbf{Q}^\top. \tag{4.22}$$

In our experience, numerical precision or some other error source (probably asymmetry in the implementation of $\mathbf{G}$ and $\mathbf{G}^\top$) leads to asymmetry in $\mathbf{K}$ that is detrimental to the

Hessian approximation. To circumvent that issue, we take several steps. First we substitute a positive definite form of $\mathbf{K}$ into Eq. 4.22,

$$\hat{\mathbf{A}} \approx \mathbf{Q} \left(\mathbf{K}\mathbf{K}^\top \right)^{\frac{1}{2}} \mathbf{Q}^\top. \tag{4.23}$$

After taking the SVD, $\mathbf{K} = \mathbf{W}\boldsymbol{\Lambda}_1\mathbf{Z}^\top$, we substitute it into Eq. 4.23 and simplify to determine a symmetric low-rank form for $\hat{\mathbf{A}}$:

$$\begin{aligned}
\hat{\mathbf{A}} &\approx \mathbf{Q} \left(\mathbf{W}\boldsymbol{\Lambda}_1\mathbf{Z}^\top\mathbf{Z}\boldsymbol{\Lambda}_1\mathbf{W}^\top \right)^{\frac{1}{2}} \mathbf{Q}^\top \\
\hat{\mathbf{A}} &\approx \mathbf{Q} \left(\mathbf{W}\boldsymbol{\Lambda}_1^2\mathbf{W}^\top \right)^{\frac{1}{2}} \mathbf{Q}^\top \\
\hat{\mathbf{A}} &\approx \mathbf{Q}\mathbf{W}\boldsymbol{\Lambda}_1\mathbf{W}^\top\mathbf{Q}^\top.
\end{aligned} \tag{4.24}$$

$\mathcal{H}_{\delta \boldsymbol{v},o}$ is transformed to $\mathcal{H}_{\delta \boldsymbol{v}}$ by adding 1 to the eigenvalues. Algorithm 3 in Appendix D summarizes the incremental 4D-Var procedure for RIOT-56.

(2) **RIOT-51**: Let $\hat{\mathbf{A}} \equiv \mathcal{H}_{\delta \boldsymbol{v},o}^{\frac{1}{2}} = \mathbf{R}^{-\frac{1}{2}}\mathbf{G}\mathbf{U} \in \mathbb{R}^{m \times n}$

In this case, we set $\hat{\mathbf{A}}$ equal to the right square-root of $\mathcal{H}_{\delta \boldsymbol{v},o}$, where $\mathbf{A} = \hat{\mathbf{A}}^\top\hat{\mathbf{A}}$. $\mathbf{Y} \in \mathbb{R}^{m \times l}$ is found by multiplying $\hat{\mathbf{A}}$ by $N_{\text{app}} = (l + p)$ random vectors, $\boldsymbol{\omega}_{k_i} \in \mathbb{R}^n$, in preconditioned CV space. Once $\mathbf{Q} \in \mathbb{R}^{m \times N_{\text{app}}}$ is known, we apply the RSVD from Algorithm 5.1 of HMT11 for non-square $\hat{\mathbf{A}}$. The next step from that algorithm is to find $\mathbf{K}_1 = \mathbf{Q}^\top\hat{\mathbf{A}}$, but we only have the ability to multiply $\hat{\mathbf{A}}$ or $\hat{\mathbf{A}}^\top$ by a vector, because they are not represented explicitly. Therefore we instead find $\mathbf{K}_1^\top = \hat{\mathbf{A}}^\top\mathbf{Q}$ and take its transpose.

Next we take the SVD, $\mathbf{K}_1 = \mathbf{W}\mathbf{S}\mathbf{Z}^\top$. HMT11 state that the wall-time of a standard SVD algorithm operating on a rank-l matrix of dimension $n \times m$ scales as $\mathcal{O}(nml)$. Since $\mathbf{K}_1 \in \mathbb{R}^{l \times n}$, this SVD should scale as $\mathcal{O}(nl^2)$. Therefore, if this SVD takes less than 10 seconds for $n = 3 \times 10^5$ using LAPACK on a single processor (which it does), it should scale conveniently for larger problems (e.g., NWP and air quality forecasting).

With $\mathbf{K}_1$ known, the observation Hessian is decomposed as

$$\mathcal{H}_{\delta \boldsymbol{v},o} = \hat{\mathbf{A}}^\top\hat{\mathbf{A}} \approx \mathbf{Z}\mathbf{S}\mathbf{W}^\top\mathbf{Q}^\top\mathbf{Q}\mathbf{W}\mathbf{S}\mathbf{Z}^\top = \mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^\top \tag{4.25}$$

The eigenvalues, $\mathbf{\Lambda} = \mathbf{S}^2$, are transformed to those of $\mathcal{H}_{\delta \boldsymbol{v}}$ by adding 1. Algorithm 4 in Appendix D summarizes the incremental 4D-Var procedure for RIOT-51.

(3) **RIOT-51$^\top$**: Let $\hat{\mathbf{A}} \equiv \mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-\frac{1}{2}} \in \mathbb{R}^{n \times m}$

The process for this form of $\hat{\mathbf{A}}$ is similar to RIOT-51 except that $\boldsymbol{\omega}_i \in \mathbb{R}^m$, $\mathbf{Y} \in \mathbb{R}^{n \times N_{\text{app}}}$, $\mathbf{Q} \in \mathbb{R}^{n \times N_{\text{app}}}$, and $\mathcal{H}_{\delta \boldsymbol{v},o} = \hat{\mathbf{A}} \hat{\mathbf{A}}^\top$.

The three different algorithmic approaches to RSVD use the same components, namely the TLM, ADM, observation covariance, and a set of random vectors. In perfect arithmetic they should all lead to the same answer for a low-rank approximation to the symmetric, positive definite $\mathcal{H}_{\delta \boldsymbol{v}}$. Following from the summary of single-pass methods by HMT11, we expect RIOT-56 to be less successful than RIOT-51 or RIOT-51$^\top$. In the event that there are scalability issues in RIOT-51 with forming the SVD of $\mathbf{K}_1$, RIOT-56 is a usable alternative. We evaluate both RIOT-56, RIOT-51, and RIOT-51$^\top$ in Sec. 4.3 to characterize their differences in application to 4D-Var.

So far we assumed that the randomized range finder, HMT11 Algorithm 4.1, is used to find the basis set, $\mathbf{Q}$, of the target matrix, $\hat{\mathbf{A}}$. Of the randomized range approximation schemes described by HMT11, this one is the easiest to implement and has the shortest wall-time. However, the basis approximation error in Eq. 4.20 can only be reduced by increasing the number of random samples, N_{app}. In incremental 4D-Var this equates to additional TLM and ADM ensemble members, which for a fixed wall-time could increase the number of processors needed beyond an acceptable limit. Since the precision of $\mathbf{Q}$ will impact the optimality of the analysis increment or posterior covariance, a range finder compatible with a fixed processor count could be advantageous, especially if the eigenvalues of the Hessian decay slowly.

An alternative is to use power iterations, e.g., HMT11 Algorithms 4.3 and 4.4. The randomized power iteration originates from Rokhlin et al. (2010) and is based on classical orthogonal iteration methods (Golub and Van Loan (1996), p. 332). The power iteration derives its name

from repeated multiplications by the target matrix and its transpose,

$$\mathbf{Y} = \left(\hat{\mathbf{A}}\hat{\mathbf{A}}^\top\right)^q \hat{\mathbf{A}}\boldsymbol{\Omega}. \tag{4.26}$$

When $q = 0$, Eq. 4.26 simplifies to Eq. 4.19. $\mathbf{Q}$ is found from QR factorization of $\mathbf{Y}$. The randomized subspace iteration (RSI), HMT11 Algorithm 4.4, uses repeated orthonormalization in order to preserve information about the weaker singular modes of $\hat{\mathbf{A}}$ in floating-point arithmetic. In contrast, the randomized power iteration, HMT11 Algorithm 4.3, does not use orthonormalization. The reorthonormalizaiton uses intermediate matrices $\widetilde{\mathbf{Y}}_j = \hat{\mathbf{A}}^\top \mathbf{Q}_{j-1}$ and $\mathbf{Y}_j = \hat{\mathbf{A}}\widetilde{\mathbf{Q}}_j$, where $\widetilde{\mathbf{Q}}_j$ is the orthonormalized $\widetilde{\mathbf{Y}}_j$ and $\mathbf{Q}_{j-1}$ is the orthonormalized $\mathbf{Y}_{j-1}$. This need for enforced orthogonality, and the power iteration in general, resembles Krylov subspace methods. When $q = 0$, both power iteration methods simplify to the randomized range finder described in the RIOT-56 or RIOT-51. Herein, we will also refer to RIOT methods with $q > 0$, indicating that the RSI is used instead of the randomized range finder.

4.2.2.3 RSVD with a hybrid basis

There are large differences between Krylov and randomized methods, and one of these is in how the first column of $\mathbf{Q}$ is determined, which is the dominant basis vector, $\boldsymbol{q}_1$. One Krylov method, CG, solves a problem, $\mathbf{A}\delta\boldsymbol{v} = \boldsymbol{b}$, where $\boldsymbol{q}_1 = \hat{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\sqrt{\boldsymbol{b}^\top \boldsymbol{b}}}$. $\hat{\boldsymbol{b}}$ is the steepest descent direction. In RSVD, the randomized range finder and subsequent decomposition are aimed solely at $\hat{\mathbf{A}}$ without regards for $\boldsymbol{b}$. In the first iteration of RIOT-56, we found that the columns of $\mathbf{Y}$ are nearly (but not perfectly) parallel or anti-parallel to $\boldsymbol{b} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-1} \boldsymbol{d}^o$. Therefore, in the incremental 4D-Var problem, the range of the Hessian shares directional information with the steepest descent direction, $\hat{\boldsymbol{b}}$. When applying RSVD to this problem it is beneficial to consider the system of equations,

$$\mathbf{A}\delta\boldsymbol{v} = \boldsymbol{b} \tag{4.27}$$

$$\hat{\mathbf{A}}\boldsymbol{\Omega} = \mathbf{Y} \tag{4.28}$$

$$\hat{\mathbf{A}}\mathbf{Q} = \mathbf{Q}\mathbf{K}, \tag{4.29}$$

where $\mathbf{Q}$ comes from QR($\mathbf{Y}$). $\mathbf{A}$ is either equal to $\hat{\mathbf{A}}$ or a product of $\hat{\mathbf{A}}$ and $\hat{\mathbf{A}}^\top$, depending on the matrix decomposition method (see Sec. 4.2.2.2). For instructive purposes we continue the discussion using the randomized range finder (i.e., $q = 0$) with RIOT-56 (i.e., $\mathbf{A} = \hat{\mathbf{A}}$). We will subsequently discuss the application to RIOT-51 and RIOT-51$^\top$ for $q > 0$.

For RIOT-56, Eq. 4.21 combines Eqs. 4.28 and 4.29. Utilizing $\mathbf{A} = \hat{\mathbf{A}}$, and the fact that $\hat{\boldsymbol{b}}$ is to $\boldsymbol{b}$ as $\boldsymbol{q}_1$ is to $\boldsymbol{y}_1$, the system of equations can be represented as

$$\hat{\mathbf{K}} \begin{bmatrix} \hat{\boldsymbol{b}} & \mathbf{Q} \end{bmatrix}^\top [\delta \boldsymbol{v} \quad \boldsymbol{\Omega}] \approx \begin{bmatrix} \hat{\boldsymbol{b}} & \mathbf{Q} \end{bmatrix}^\top [\boldsymbol{b} \quad \mathbf{Y}], \tag{4.30}$$

where $\hat{\mathbf{K}} \in \mathbb{R}^{N_{\text{app}}+1 \times N_{\text{app}}+1}$ is a reduced form of $\mathbf{A}$ that also accounts for the least squares problem at hand. We can also define $\hat{\mathbf{Y}} = [\boldsymbol{b} \quad \mathbf{Y}] \in \mathbb{R}^{n \times N_{\text{app}}+1}$, and if $\mathbf{q}_i \perp \hat{\boldsymbol{b}}$ for all the columns of $\mathbf{Q}$, then we have $\hat{\mathbf{Q}} = \begin{bmatrix} \hat{\boldsymbol{b}} & \mathbf{Q} \end{bmatrix}$. The new problem is then

$$\hat{\mathbf{K}}\hat{\mathbf{Q}}^\top [\delta \boldsymbol{v} \quad \boldsymbol{\Omega}] \approx \hat{\mathbf{Q}}^\top \hat{\mathbf{Y}}, \tag{4.31}$$

and this expands to the separable system of equations,

$$\hat{\mathbf{K}}\hat{\mathbf{Q}}^\top \boldsymbol{\Omega} \approx \hat{\mathbf{Q}}^\top \mathbf{Y} \tag{4.32}$$

$$\mathbf{A}\delta \boldsymbol{v} \approx \hat{\mathbf{Q}}\hat{\mathbf{K}}\hat{\mathbf{Q}}^\top \delta \boldsymbol{v} = \boldsymbol{b}. \tag{4.33}$$

Introducing a single new basis vector, $\hat{\boldsymbol{b}}$, may affect the approximation error of $\mathbf{A} \approx \mathbf{Q}\hat{\mathbf{K}}\mathbf{Q}^\top$, but it can never reduce that error below that of the rank-l truncated SVD. However, including $\hat{\boldsymbol{b}}$ does change the analysis increment when a randomized power iteration is used, which we show in Sec. 4.3. Eq. 4.26 requires multiplication by both $\hat{\mathbf{A}}$ and its transpose, which for RIOT-56 is the full observation Hessian (i.e., $\hat{\mathbf{A}}^\top = \hat{\mathbf{A}} = \mathcal{H}_{\delta \boldsymbol{v},o}$). We forgo a rigorous theoretical extension to RIOT-51 and RIOT-51$^\top$, but infer that simply appending $\hat{\boldsymbol{b}}$ onto $\mathbf{Q}$ as the leading basis vector, $\boldsymbol{q}_1$, is equivalent to solving for an increment that simultaneously solves the least-squares and RSVD problems. Algorithms 1 and 2 describe the hybrid procedures for RIOT-51 and RIOT-51$^\top$, respectively. The hybridization takes place in step 10 of Algorithm 1 and steps 5 and 15 of Algorithm 2. Without the prepending by $\hat{\boldsymbol{b}}$ in these two steps, the algorithms are equivalent to using RIOT-51 and RIOT-51$^\top$ with the range finder replaced by RSI.

Algorithm 1 Hybrid RIOT-51

Require: $\hat{\mathbf{A}} \equiv \mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o} = \mathbf{R}^{-\frac{1}{2}}\mathbf{G}\mathbf{U} \in \mathbb{R}^{m\times n}$
and $\boldsymbol{\Omega} \in \mathbb{R}^{n\times N_{\text{app}}} \sim \mathcal{N}(0,1)$
Start with $x^{k=0} = x^b$, $\boldsymbol{v}^0 = \mathbf{0}$
for $k = 1, 2, \ldots, k_f$ **do**
  Steps 3 to 12 of Algorithm 4
  Let $\mathbf{Y}_0 = \mathbf{Y}$ and find $\mathbf{Q}_0 \in \mathbb{R}^{m\times N_{\text{app}}}$ from $\text{QR}(\mathbf{Y}_0)$
  if $q > 0$ **then**
    Let $\hat{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\sqrt{\boldsymbol{b}^\top \boldsymbol{b}}}$
    for $j = 1, 2, \ldots, q$ **do**
      for all $k_i \in \{1, 2, \ldots, N_{\text{app}}\}$ **do in parallel**
        $\widetilde{\boldsymbol{y}}_{k_i,j} = \hat{\mathbf{A}}^\top \boldsymbol{q}_{k_i,j-1}$
      end for
      Calculate $\widetilde{\mathbf{Q}}_j \in \mathbb{R}^{n\times N_{\text{app}}}$ from $\text{QR}(\widetilde{\mathbf{Y}}_j)$
      Let $\hat{\mathbf{Q}}_j = \left[\hat{\boldsymbol{b}} \quad \widetilde{\mathbf{Q}}_j\,(:, 1 : N_{\text{app}} - 1)\right]$
      for all $k_i \in \{1, 2, \ldots, N_{\text{app}}\}$ **do in parallel**
        $\boldsymbol{y}_{k_i,j} = \hat{\mathbf{A}}\hat{\boldsymbol{q}}_{k_i,j}$
      end for
      Calculate $\mathbf{Q}_j \in \mathbb{R}^{m\times N_{\text{app}}}$ from $\text{QR}(\mathbf{Y}_j)$
    end for
  end if
  $\mathbf{Q} = \mathbf{Q}_q$
  Steps 14 to 22 of Algorithm 4
end for
$\mathbf{P}^a = \mathbf{U}\mathbf{P}^a_{\boldsymbol{v}}\mathbf{U}^\top$

In the incremental 4D-Var application, $\boldsymbol{b} \in \mathbb{R}^n$ is always in the control vector space. Thus, $\hat{\boldsymbol{b}}$ can only be prepended to a matrix $\mathbf{Q} \in \mathbb{R}^{n\times N_{app}}$. When the RSI is used with RIOT-51, that size restriction is met by $\widetilde{\mathbf{Q}}$, while for RIOT-51$^\top$, it is met by $\mathbf{Q}$. After several iterations of the RSI, one can imagine that the leading basis vector would no longer be equal to $\hat{\boldsymbol{b}}$. Since that would violate Eqs. 4.32 and 4.33, the basis vector $\hat{\boldsymbol{b}}$ is prepended to either $\mathbf{Q}$ or $\widetilde{\mathbf{Q}}$ during each iteration of RSI, and the last column is truncated to maintain a constant width ensemble. The repeated multiplications of $\hat{\boldsymbol{b}}$ by $\hat{\mathbf{A}}^\top$ and $\hat{\mathbf{A}}$ are identical to Krylov methods that find a basis for the set of vectors made up of increasing powers of $\mathbf{A}$ multiplied by $\hat{\boldsymbol{b}}$. Thus, we now have a method that combines the properties of Krylov and RSVD methods and uses a hybridized deterministic and stochastic basis $\hat{\mathbf{Q}}$

to reduce the number of sequential iterations and/or ensembles required to converge on a solution to the least-squares problem. This approach can scale to available computational resources of time or processor count.

Algorithm 2 Hybrid RIOT-51$^\top$

Require: $\hat{\mathbf{A}} \equiv \mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-\frac{1}{2}} \in \mathbb{R}^{n \times m}$
and $\boldsymbol{\Omega} \in \mathbb{R}^{m \times N_{\text{app}}} \sim \mathcal{N}(0,1)$

Start with $x^{k=0} = x^b$, $\boldsymbol{v}^0 = \mathbf{0}$
for $k = 1, 2, \ldots, k_f$ **do**
  Steps 3 to 12 of Algorithm 4
  Let $\mathbf{Y}_0 = \mathbf{Y}$ and find $\mathbf{Q}_0 \in \mathbb{R}^{n \times N_{\text{app}}}$ from QR($\mathbf{Y}_0$)
  Let $\hat{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\sqrt{\boldsymbol{b}^\top \boldsymbol{b}}}$
  Let $\hat{\mathbf{Q}}_0 = \begin{bmatrix} \hat{\boldsymbol{b}} & \mathbf{Q}_0 \left(:, 1 : N_{\text{app}} - 1\right) \end{bmatrix}$
  if $q > 0$ **then**
    for $j = 1, 2, \ldots, q$ **do**
      for all $k_i \in \{1, 2, \ldots, N_{\text{app}}\}$ **do in parallel**
        $\widetilde{\boldsymbol{y}}_{k_i, j} = \hat{\mathbf{A}}^\top \hat{\boldsymbol{q}}_{k_i, j-1}$
      end for
      Calculate $\widetilde{\mathbf{Q}}_j \in \mathbb{R}^{m \times N_{\text{app}}}$ from QR($\widetilde{\mathbf{Y}}_j$)
      for all $k_i \in \{1, 2, \ldots, N_{\text{app}}\}$ **do in parallel**
        $\boldsymbol{y}_{k_i, j} = \hat{\mathbf{A}} \widetilde{\boldsymbol{q}}_{k_i, j}$
      end for
      Calculate $\mathbf{Q}_j \in \mathbb{R}^{n \times N_{\text{app}}}$ from QR($\mathbf{Y}_j$)
      Let $\hat{\mathbf{Q}}_j = \begin{bmatrix} \hat{\boldsymbol{b}} & \mathbf{Q}_j \left(:, 1 : N_{\text{app}} - 1\right) \end{bmatrix}$
    end for
  end if
  $\mathbf{Q} = \hat{\mathbf{Q}}_q$
  Steps 14 to 18 of Algorithm 4
  Form eigenmodes of $\mathcal{H}_{\delta \boldsymbol{v}}$: $\boldsymbol{\Lambda} = \mathbf{S}^2 + \mathbf{I}$ and $\mathbf{V} = \mathbf{Q}\mathbf{W}$
  Steps 20 to 22 of Algorithm 4
end for
$\mathbf{P}^a = \mathbf{U}\mathbf{P}^a_{\boldsymbol{v}}\mathbf{U}^\top$

4.2.3 Evaluation Metrics

In this section, we derive two evaluation metrics that can be used to objectively compare different Hessian approximation methods. Both of the metrics require having an exact evaluation of either the left or right square-root of the Hessian.

4.2.3.1 Exact Hessian

The TLM and ADM are able to evaluate multiplications of the model Jacobian or its adjoint by vectors of size n or m, respectively. When the problem size is small enough, either $m \lesssim 10^2$ or $n \lesssim 10^2$ for the relatively expensive model evaluations of atmospheric problems, a square-root of the exact linearized Hessian (within machine precision) can be determined using an ensemble of TLM or ADM integrations at moderate cost. This also requires knowing $\mathbf{R}^{-\frac{1}{2}}$, which is straightforward in our test scenarios where $\mathbf{R}$ is assumed to be diagonal. In this work, we use a problem where m is small enough to calculate the left square-root (LSR) of the observation Hessian, explicitly:

$$\mathcal{H}_{\delta \boldsymbol{v},o,LSR} = \mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-\frac{1}{2}} \mathbf{I}_m. \tag{4.34}$$

The multiplication by $\mathbf{I}_m$ indicates that we independently evaluate each matrix-vector product with the columns of the identity matrix. Just like for RSVD, the matrix-matrix multiplication is done in parallel. Once $\mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o}$ is known, its SVD is taken as $\mathcal{H}_{\delta \boldsymbol{v},o,LSR} = \mathbf{WSZ}^\top$, and the observation Hessian is expressed as

$$\mathcal{H}_{\delta \boldsymbol{v},o} = \mathbf{WSZ}^\top \mathbf{ZSW}^\top = \mathbf{W\Lambda W}^\top, \tag{4.35}$$

where $\mathbf{\Lambda} = \mathbf{S}^2$. Once again $\mathcal{H}_{\delta \boldsymbol{v}}$ is found by adding 1 to the eigenvalues.

4.2.3.2 Range approximation error

The norm declared in Eq. 4.20 indicates there is an upper limit on approximation error, ϵ, for any low-rank basis, whether it is generated through the Lanczos recurrence or RSVD. The difficulty is in evaluating this norm when n is very large, since $\mathbf{A} = \mathcal{H}_{\delta \boldsymbol{v},o}$ can not be stored in memory, and the number of computations is large. After carrying out the exact Hessian approximation, we have an SVD,

$$\mathcal{H}_{\delta \boldsymbol{v},o,LSR} = \mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-\frac{1}{2}} = \mathbf{WSZ}^\top, \tag{4.36}$$

where $\mathcal{H}_{\delta \boldsymbol{v},o,LSR}$ is the LSR of the observation Hessian. Substituting the SVD into Eq. 4.20, we find that the spectral norm, which we will call ϵ_Q, can be transformed to

$$\epsilon_Q = \| \left[\mathbf{I}_n - \mathbf{Q}\mathbf{Q}^\top \right] \mathcal{H}_{\delta \boldsymbol{v},o} \|_2 = \| \left[\mathbf{W}\mathbf{S}\mathbf{Z}^\top - \mathbf{Q}\mathbf{Q}^\top \mathbf{W}\mathbf{S}\mathbf{Z}^\top \right] \mathbf{Z}\mathbf{S}\mathbf{W}^\top \|_2. \tag{4.37}$$

If $\mathbf{Q}$ is a full-rank basis for $\mathcal{H}_{\delta \boldsymbol{v},o}$, then $\epsilon_Q = 0$ in exact arithmetic. When $\mathbf{Q} = \mathbf{W}_l$, where $\mathbf{W}_l$ contains the leading l left-singular vectors of $\mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-\frac{1}{2}}$ (equivalent to leading eigenvectors of $\mathcal{H}_{\delta \boldsymbol{v},o}$ and $\mathcal{H}_{\delta \boldsymbol{v}}$), then $\epsilon_Q = \lambda_{l+1}$, which corresponds to the best possible rank-l approximation of the Hessian. The calculation of ϵ_Q can be further simplified by taking the SVD of the term in brackets on the right-hand-side of Eq. 4.37,

$$\left[\mathbf{W}\mathbf{S}\mathbf{Z}^\top - \mathbf{Q}\mathbf{Q}^\top \mathbf{W}\mathbf{S}\mathbf{Z}^\top \right] = \left(\mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1^\top \right) \in \mathbb{R}^{n \times m}. \tag{4.38}$$

ϵ_Q then simplifies to

$$\epsilon_Q = \| \mathbf{U}_1 \mathbf{S}_1 \mathbf{V}_1^\top \mathbf{Z}\mathbf{S}\mathbf{W}^\top \|_2. \tag{4.39}$$

Then take the SVD of

$$\mathbf{S}_1 \mathbf{V}_1^\top \mathbf{Z}\mathbf{S} = \left(\mathbf{U}_2 \mathbf{S}_2 \mathbf{V}_2^\top \right) \in \mathbb{R}^{l \times m}, \tag{4.40}$$

which gives

$$\epsilon_Q = \| \mathbf{U}_1 \mathbf{U}_2 \mathbf{S}_2 \mathbf{V}_2^\top \mathbf{W}^\top \|_2, \tag{4.41}$$

and finally

$$\epsilon_Q = \max \mathbf{S}_2. \tag{4.42}$$

With this method to find ϵ_Q, we can now objectively compare different approximations of $\mathbf{Q}$. The SVD of $\mathcal{H}_{\delta \boldsymbol{v},o,LSR}$ was found in MATLAB using the "svd" function.

4.2.3.3 Bayes risk

The Bayes risk is a wholistic metric that can be used to compare the analysis increments associated with each Hessian approximation method. It is defined as

$$\| \boldsymbol{x} - \boldsymbol{x}^a \|^2_{(\mathbf{P}^a)^{-1}} = \left(\boldsymbol{x} - \boldsymbol{x}^a \right)^\top \left(\mathbf{P}^a \right)^{-1} \left(\boldsymbol{x} - \boldsymbol{x}^a \right). \tag{4.43}$$

As shown in Eqs. 4.14 and 4.15, BH16 gave theoretical expectancy results for the Bayes risk associated with the rank-l LRU and LRA approximations of the Hessian. Since the Lanczos recurrence and RSVD do not produce identical increments to a truncated SVD, it would be useful to be able to evaluate this same metric given a preconditioned analysis increment, $\delta\boldsymbol{v}$, from either method and the true increment, $\delta\boldsymbol{v}^a$. $\delta\boldsymbol{v}^a$ is calculated using the full-rank inverse Hessian SVD from Sec. 4.2.3.1.

When the preconditioned inverse posterior covariance, $\left(\mathbf{P}_{\boldsymbol{v}}^a\right)^{-1} = \mathcal{H}_{\delta\boldsymbol{v}}$, is combined with Eq. 3.25, we find that

$$\left(\mathbf{P}_{\boldsymbol{v}}^a\right)^{-1} = \mathcal{H}_{\delta\boldsymbol{v}} = \mathbf{U}^\top \left(\mathbf{P}^a\right)^{-1} \mathbf{U}. \tag{4.44}$$

Premultiplying by $\mathbf{B}^{-1}\mathbf{U}$ and postmultiplying by $\mathbf{U}^\top\mathbf{B}^{-1}$, the posterior covariance is equal to

$$\left(\mathbf{P}^a\right)^{-1} = \mathbf{B}^{-1}\mathbf{U}\mathcal{H}_{\delta\boldsymbol{v}}\mathbf{U}^\top\mathbf{B}^{-1}. \tag{4.45}$$

The difference between the posterior and the truth can also be simplified to

$$\boldsymbol{x} - \boldsymbol{x}^a = \mathbf{U}\left(\boldsymbol{v} - \boldsymbol{v}^a\right) = \mathbf{U}\left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right). \tag{4.46}$$

Substituting Eqs. 4.45 and 4.46 into Eq. 4.43, the Bayes risk is

$$\begin{aligned} \left\|\boldsymbol{x} - \boldsymbol{x}^a\right\|^2_{\left(\mathbf{P}^a\right)^{-1}} &= \left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right)^\top \mathbf{U}^\top\mathbf{B}^{-1}\mathbf{U}\mathcal{H}_{\delta\boldsymbol{v}}\mathbf{U}^\top\mathbf{B}^{-1}\mathbf{U}\left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right) \\ &= \left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right)^\top \mathcal{H}_{\delta\boldsymbol{v}}\left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right) \end{aligned} \tag{4.47}$$

Using the eigendecomposition from Eq. 4.35, this simplifies to

$$\left\|\boldsymbol{x} - \boldsymbol{x}^a\right\|^2_{\left(\mathbf{P}^a\right)^{-1}} = \left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right)^\top \mathbf{W}\left(\boldsymbol{\Lambda} + \mathbf{I}\right)\mathbf{W}^\top\left(\delta\boldsymbol{v} - \delta\boldsymbol{v}^a\right). \tag{4.48}$$

4.3 Results

There are four types of approximation errors we must account for when considering a particular Newton-based nonlinear optimization algorithm:

(1) The GN algorithm introduces error by linearizing the cost function. We mitigate that error through DGN, and through successive relinearization of the the problem around new states.

(2) When we assume that the Hessian can be approximated by its dominant eigenmodes, we are discarding some information. The appropriate rank must balance the needs to resolve fine scale information that is useful to the optimization, to avoid overfitting unrealistic modes, and to utilize available computational resources.

(3) For a particular inner-loop method (Lanczos recurrence, RSVD), how close do l iterations or ensembles come to being equivalent to a rank-l approximation of the Hessian? This error source reflects the need for the oversampling parameter, p.

(4) Is the LRA or the LRU utilized, and what difference does it make?

The error, ϵ, from Eq. 4.20 is contained in items (2) and (3) above. We will explore the impacts of each of these error sources through demonstrations on a problem with small m using the Hessian approximation methods described in Sec. 4.2.

The test problem is described in detail in Chapter 3. We use aircraft observations of black carbon (BC) throughout California, measured with a single particle soot photometer (SP2) (Sahu et al., 2012) on board a DC-8 aircraft during the Arctic Research of the Composition of the Troposphere from Aircraft and Satellites in collaboration with the California Air Resources Board (ARCTAS-CARB) field campaign (Jacob et al., 2010). For the analysis herein, we focus on the flight of 22 June, 2008 which characterized local emissions of trace gases and aerosols from anthropogenic and biomass burning (BB) sources as well as inflow from the Pacific Ocean. On that day, there were $m = 241$ aircraft observations. A large portion of the flight time was spent characterizing the Los Angeles boundary layer between 8:00 and 10:00 LT. Most of the remainder of the flight was spent flying out over the ocean, up to Crescent City, and then returning down the coast to Los Angeles. In Chapter 3, we assessed the utility of these observations in constraining 18 km^2 gridded sources of BC from both anthropogenic and BB activity. Herein, the purpose is to compare computational and accuracy metrics for the different methods of Hessian approximation and posterior updates for this specific problem. The CV, $\boldsymbol{x}$ is comprised of the exponential emission scaling factors for 79×79 grid cells and 24 hourly emission bins. Since the domain includes

the ocean and the fires are sparse, many of the grid cells ($> \frac{1}{2}$) do not contain emissions for one or both sectors. So although $n = 299,568$, and the matrices we consider will have dimensions of n and m, the effective problem size is $n_{\text{eff}} = 121,912$.

4.3.1 First outer iteration

Before assessing the global convergence properties of a Lanczos-GN optimization and RIOT across multiple outer iterations, it is instructive to start with a single outer iteration. For this purpose, we apply the Lanczos recurrence, RSVD, and the exact Hessian method within the first outer iteration of 4D-Var for the 22 June scenario. Over the course of this work, we found that WRFDA-Chem does not have a perfectly symmetric preconditioned observation Hessian $\mathcal{H}_{\delta \boldsymbol{v},o}$. This is due to inconsistencies in the action on a vector between the TLM operator ($\mathbf{G}\delta\boldsymbol{x}$) and ADM operator ($\mathbf{G}^\top\delta\boldsymbol{y}$) or between the preconditioner ($\mathbf{U}\delta\boldsymbol{v}$) and its transpose ($\mathbf{U}^\top\delta\boldsymbol{x}$). All of the methods we are applying are meant for symmetric matrices, and the underlying theory of this problem states that the Hessian should be symmetric.

Fortunately there is an objective way to evaluate the matrix approximation methods without diagnosing the symmetry problem. We start by evaluating the LSR of $\mathcal{H}_{\delta \boldsymbol{v},o}$ using the adjoint model as described in Sec. 4.2.3.1. Then we form a perfectly symmetric analogue of the Hessian by multiplying,

$$\mathcal{H}_{\delta \boldsymbol{v},o,SYMM} = \mathcal{H}_{\delta \boldsymbol{v},o,LSR}\mathcal{H}^\top_{\delta \boldsymbol{v},o,LSR}, \tag{4.49}$$

where $\mathcal{H}_{\delta \boldsymbol{v},o,LSR} = \mathbf{U}^\top\mathbf{G}^\top\mathbf{R}^{-\frac{1}{2}}$. We apply the Lanczos recurrence, RIOT-56, RIOT-51, Hybrid RIOT-51, and Hybrid RIOT-51$^\top$ to the problem $\mathbf{A} \equiv \mathbf{I}_n + \mathcal{H}_{\delta \boldsymbol{v},o,SYMM}$.

Figure 4.1 shows the resulting eigenspectra for the Lanczos recurrence, RIOT-56, and RIOT-51, each of which converges toward the true eigenspectrum from below. Eigenvalues from RIOT-51 and RIOT-56 match to within 10 digits. The Lanczos recurrence exhibits long tails at the end of each approximate spectrum, indicating that the trailing eigenvalues between l and the number of approximate modes, $N_{\text{app}} = l + p$, are poorly estimated relative to the others. The tails are smaller

for RIOT-51/56.

Figure 4.2 shows the absolute error between the eigenvalues of the exact Hessian and those from three approximation methods. The Lanczos recurrence quickly converges to nearly machine precision for the leading eigenvalues, with larger errors for the smallest 30-50% of eigenvalues. The values for RIOT-51 when $q = 0$ should be identical to those from RIOT-56. When $q = 0$, the error exhibits constant absolute accuracy but a stable decay in relative accuracy from the largest to the smallest eigenvalues. Increasing q substantially reduces error in all eigenvalues, especially the leading ones. This is consistent with the theory discussed by HMT11. Hybridization does not have much impact on the eigen spectra for RIOT-51. Also, RIOT does not meet or exceed the precision of the Lanczos recurrence until the final 20% of the spectrum, but shows marked improvement with $q > 0$.

Figure 4.3(a) shows ϵ_Q for the truncated SVD and for several approximate bases. As predicted by theory, ϵ_Q lines up very well with λ_{l+1} when $\mathbf{Q}$ is equal to the left-singular-vectors of the truncated SVD ($p = 0$). The other two lines in Fig. 4.3(a) correspond to the basis from the Lanczos recurrence and the expected basis from 20 realizations of RIOT-56. Figure 4.3(d) includes ϵ_Q for a single realization of RIOT-51 with the RSI and multiple values of q. When $q = 0$, RSI collapses to the randomized range finder in the standard RIOT-51. We have found that with a perfectly symmetric $\hat{\mathbf{A}}$, the basis and eigenmodes produced from RIOT-51 and RIOT-56 with the randomized range finder are identical for a given realization of $\mathbf{\Omega}$. For any $q > 0$, the spectral norm is significantly reduced to values lower than those produced by the Lanczos recurrence, eventually converging to the truncated SVD. This is consistent with the convergence of the power iteration discussed by HMT11.

When a horizontal line is drawn from any of the colored lines in Figs. 4.3(a,d) to the black line for the exact symmetric Hessian, the new x-value is equivalent to the effective rank, l, of the approximation method. These rank values are plotted in Figs. 4.3(b,e). The effective oversampling, p, is the vertical distance between each of the colored lines and the one-to-one line. For larger ensemble and iteration counts, the effective rank becomes more deficient for the Lanczos recurrence,

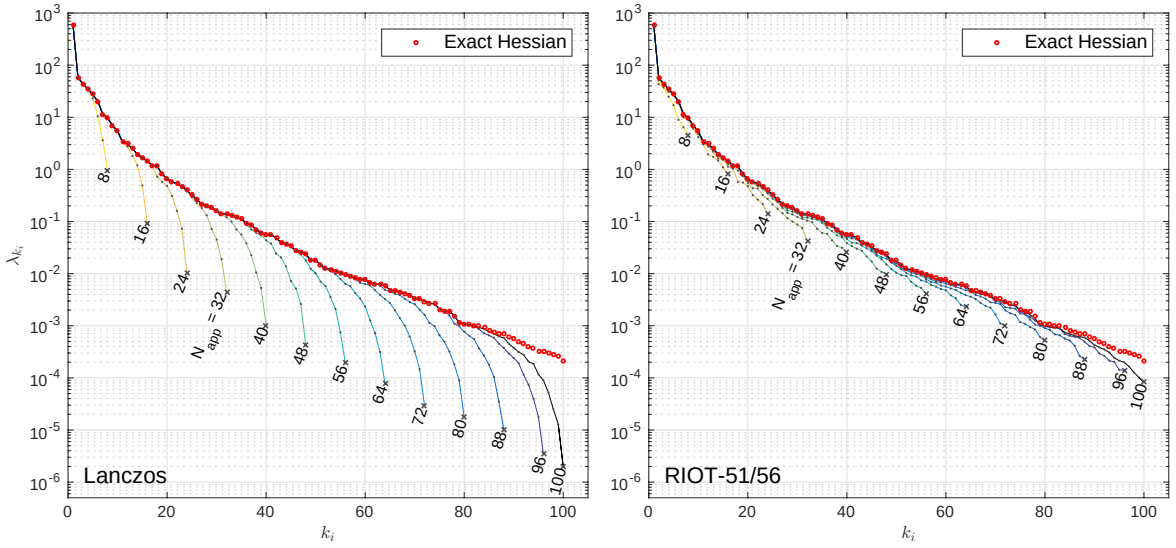

Figure 4.1: Approximate eigenvalue spectra of the perfectly symmetric observation Hessian, $\mathcal{H}_{\delta \boldsymbol{v},o,SYMM}$, for the Lanczos recurrence and RIOT-51/56 in the first outer iteration. The eigenvalues from RIOT-51 and RIOT-56 are nearly identical. Each colored line shows the estimate of the spectrum $[\lambda_1, \ldots, \lambda_{k_i=l}]$ for every eighth value of N_{app}. The black numbers on the plot are equal to the number of iterations or ensembles in a given method, N_{app}. The exact eigenvalues of the perfectly symmetric Hessian are also shown.

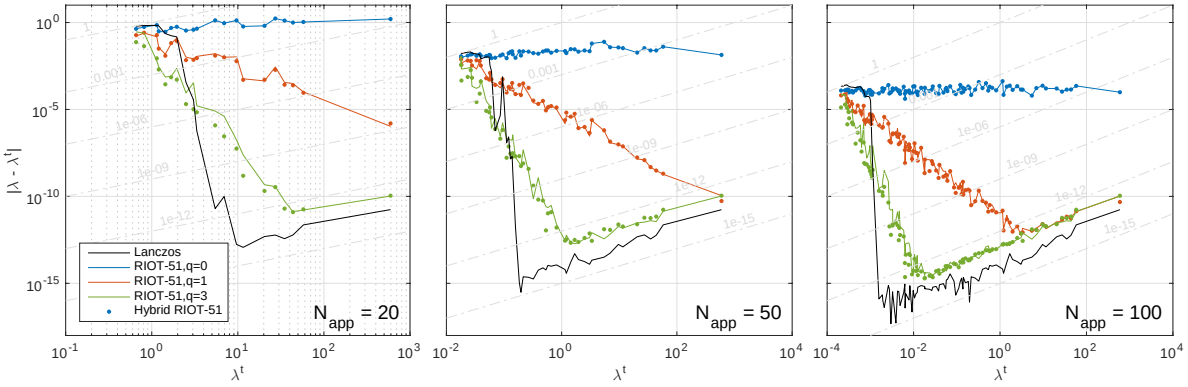

Figure 4.2: Each plot shows the absolute error between the eigenvalues of the exact Hessian and those from the Lanczos recurrence, RIOT-51 with RSI, and hybrid RIOT-51 with RSI for three different numbers of ensembles or iterations, N_{app}, and for different numbers of RSI iterations, q. The exact eigenvalue is on the x-axis. Also plotted are lines of constant fractional error between 10^{-15} and 1

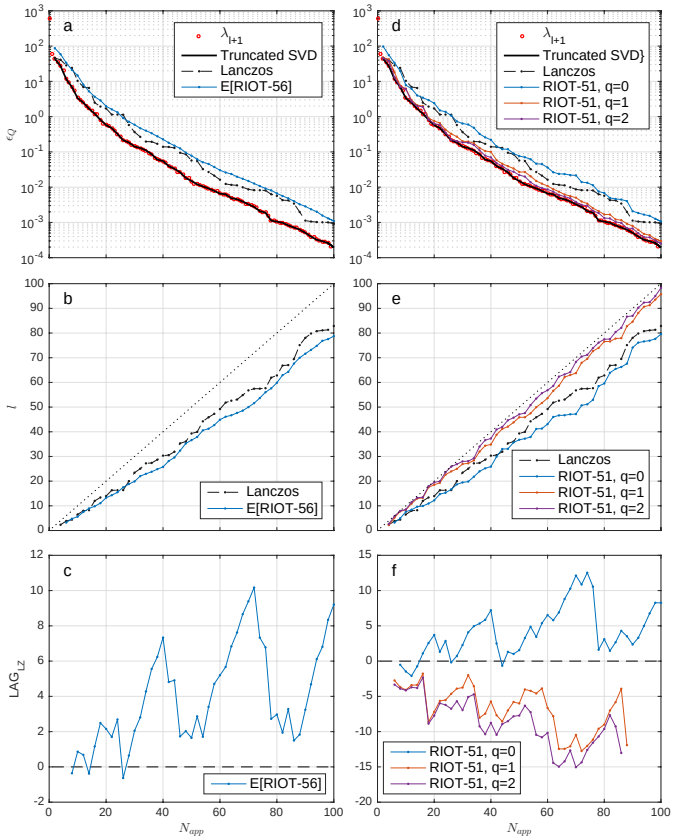

Figure 4.3: (a,d) ϵ_Q for bases formed from (1) direct SVD, (2) $\mathbf{Q}$ from the Lanczos recurrence, (3) the expectation of $\mathbf{Q}$ from RIOT-56, and (4) $\mathbf{Z}$ from a single realization of the non-hybrid RIOT-51 with RSI and $q = [0, 1, 2]$; all methods are applied to an exactly symmetric observation Hessian, $\mathcal{H}_{\delta \boldsymbol{v},o,SYMM}$. (b,e) The effective rank, l, versus number of iterations/ensembles for each approximation method, and (c,f) the lag in rank between RIOT methods and the Lanczos recurrence baseline versus the number of iterations/ensembles, $N_{\text{app}} = l + p$.

RIOT-56, and RIOT-51 when q=0. Thus, as the desired rank, l, increases, more oversampling is required and p should not be assumed constant. This deficiency is much less evident for $q > 0$.

Knowing the effective rank for each method enables the calculation of $\mathrm{LAG_{LZ}}$, or the rank lag with respect to the Lanczos recurrence, shown in Figs. 4.3(c,f). $\mathrm{LAG_{LZ}}$ indicates at specific numbers of ensembles how many fewer iterations it took the Lanczos recurrence to reach the same rank. $\mathrm{LAG_{LZ}}$ is found by drawing a horizontal line from the RIOT methods in Figs. 4.3(b,e) back to the line for the Lanczos recurrence, and then subtracting the N_{app} value on the x-axis from that of RIOT. A negative lag indicates that RIOT required fewer ensemble members than the Lanczos recurrence required iterations. For RIOT-56 and RIOT-51 ($q = 0$), this occurs randomly for small N_{app}. $\mathrm{LAG_{LZ}}$ never exceeds 12 for these two cases. When $q > 0$, RIOT can produce bases that are significantly more accurate than the Lanczos recurrence at given values of N_{app}. These are not exactly one-to-one comparisons since RIOT-51 takes $\times 2$ as long when $q = 1$ and $\times 3$ as long when $q = 2$.

The lag is never known a priori, but characterizing a particular problem (i.e., BC source inversion) ahead of time gives a rough idea of how many ensembles are required in RIOT to match the effective basis rank of an equivalent implementation of the Lanczos recurrence. If a particular application uses less than 20 Lanczos iterations, then five extra ensembles ought to be more than enough to achieve a similar effective rank. With the parallelization in RIOT, these extra ensemble members add some computational cost, but they add very little to the wall-time of the inversion in the form of gathering/distributing data. In this application, the ADM and TLM simulations account for a bulk of the wall-time.

Using the LRU and LRA in Eqs. 4.11 and 4.12 and the preconditioned update formula in Eq. 4.9 we can calculate $\delta \boldsymbol{v}$ for all of the different methods, and use the truncated SVD of the exact symmetric Hessian to calculate $\delta \boldsymbol{v}^a$. Combined with the eigendecomposition of the symmetric Hessian, we can calculate the Bayes risk of each method using Eq. 4.48. Figure 4.4 shows the Bayes risk for multiple approximation methods. We focus first on Figure 4.4(a).

The Bayes risk from the truncated SVD of the exact Hessian matches very well with the

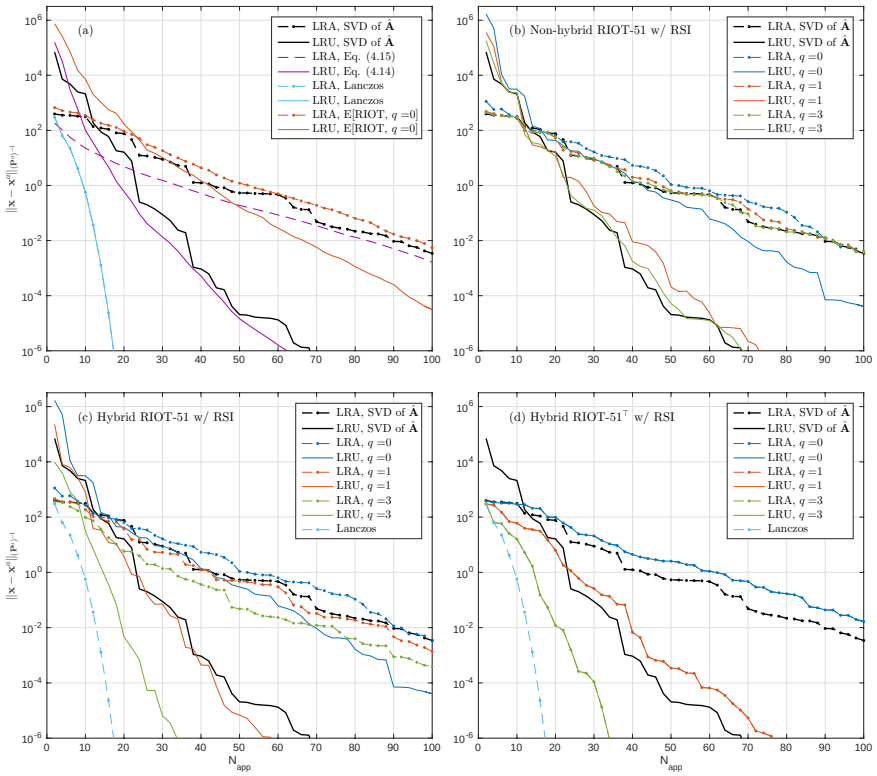

Figure 4.4: (a) Bayes risk for the truncated SVD, theoretical limits in Eqs. 4.14 and 4.15, the Lanczos recurrence, and the expectation from RIOT with $q = 0$. Values are shown for increments that use either the LRA or LRU form of the approximate Hessian. (b-d) Same as (a), but for three different variants of RIOT with varying numbers of RSI iterations, q. For RIOT-51$^\top$ in (d) and the Lanczos recurrence in (a,c,d), the LRU and LRA values coincide across all N_{app}

theoretical formulas derived by BH16, Eqs. 4.14 and 4.15. Although we do not show it here, the agreement improves steadily for higher numbers of modes($N_{\text{app}} < 75$). The Bayes risk for the LRA and LRU with the exact SVD intersect near $N_{\text{app}} = l = 14$, which is slightly less than $l = 19$, where an eigenvalue of $\mathcal{H}_{\delta \boldsymbol{v}}$ first descends below 2; the theoretical Bayes risks for the LRU and LRA intersect at the same rank, $l = 14$.

The expected increment from RIOT has a delayed intersection point between the LRA and LRU Bayes risks, but nonetheless demonstrates that the LRA is superior when only a few modes are available. As expected, RIOT is not as accurate as the truncated SVD; however, the latter requires as many ensemble members as there are observations, which is prohibitive in operational forecasting. The LRA and LRU with the Lanczos recurrence produce practically identical increments. Additionally, these increments exhibit much faster reduction in the norm than any other approach, including the truncated exact eigenmodes. This seems counterintuitive since the Lanczos recurrence produces a basis and eigenvalues that are inferior to the truncated SVD according to Figs. 4.2 and 4.3. In Sec. 4.2.2.3, we hypothesized that this behavior is attributable to the combined use of a power iteration and a simultaneous solution of the least squares and matrix decomposition problems. To evaluate this, we analyze the behavior of the hybrid RIOT methods.

Figure 4.4(b) shows the Bayes risk for a single realization of RIOT-51 with RSI and $q = [0, 1, 3]$, but without hybridization. As might be expected, the improved basis (see Fig. 4.3) achieved with higher RSI powers, q, reduces the norm for both LRA and LRU toward their respective limits defined by the truncated SVD. For the few locations where the Bayes risk of the non-hybrid RIOT-51 is lower than that of the truncated SVD, it is easily explained by the stochastic behavior of a single RIOT realization. The expectation across multiple realizations was not calculated for values of $q > 0$, due to the computational expense. We see in Fig. 4.4(c) that as q is increased for the hybridized RIOT-51 with either the LRA or LRU, the Bayes risk converges to values that are below the limits posed by the truncated SVD. In Fig. 4.4(d) the hybridized RIOT-51$^\top$ exhibits similar convergence. Additionally, the LRU and LRA give identical analysis increments for RIOT-51$^\top$, similar to the behavior of the Lanczos recurrence. These results for non-hybrid and hybrid

RIOT methods are compatible with the aforementioned hypothesis about the Lanczos recurrence. Principally, the levels of increment accuracy found with the Lanczos recurrence are attributable to the combined effects of a power iteration with $q > 2$ and a dominant basis vector equal to $\hat{\boldsymbol{b}}$.

Then the question remains, why do the LRU and LRA increments match for both the Lanczos recurrence and RIOT-51$^\top$? The hybridization step prepends $\hat{\boldsymbol{b}}$ to the CV basis, and in hybrid RIOT-51 that step is once-removed – by way of multiplication by the $\hat{\mathbf{A}}$ operator – from the final basis in observation space. That is not the case for RIOT-51$^\top$. Having $\hat{\boldsymbol{b}}$ in the final basis is the commonality between the Lanczos recurrence and hybrid RIOT-51$^\top$. When the LRU is used, the analysis increment is calculated as a multiplication of Eq. 4.11 by $\boldsymbol{b} = \frac{\partial J}{\partial \boldsymbol{v}}|_{\boldsymbol{v}_0}$, i.e.,

$$\delta \boldsymbol{v} = \left[\mathbf{I} - \sum_{k_i=1}^{N_{\text{app}}} \left(\frac{\lambda_{k_i} - 1}{\lambda_{k_i}} \right) \hat{\boldsymbol{\nu}}_{k_i} \hat{\boldsymbol{\nu}}_{k_i}^\top \right] \boldsymbol{b}. \tag{4.50}$$

Multiplying the identity matrix by the gradient and then subtracting the complement of each eigenmode admits that $\boldsymbol{b}$ is the most dominant component of the Hessian and that each other mode introduces extra information on top of it. When $\hat{\boldsymbol{b}}$ is assumed to be the dominant basis vector of the Hessian, as it is within the Lanczos recurrence and hybrid RIOT-51$^\top$, then its complement is of zero length relative to the direction of the vector by which it is multiplied, $\boldsymbol{b}$. Thus, in those two methods, the differences between LRA and LRU are superficial. Including $\hat{\boldsymbol{b}}$ as the dominant basis vector gives more efficient increments in terms of Bayes risk, because all other basis vectors are calculated as orthogonal to it. There is less information remaining about the least squares solution to resolve with the weaker modes. When the LRU is used without $\hat{\boldsymbol{b}}$ in the basis, the eigenvectors can contain redundant directional information relative to $\mathbf{I}\boldsymbol{b}$.

Since we will eventually apply DGN between the outer iterations, it is more important to match the direction of $\delta \boldsymbol{v}$ with its true value than it is to match the magnitude of $\delta \boldsymbol{v}$. Thus, the Bayes risk in Figure 4.4 is more applicable to linear problems than nonlinear problems. Not only does Bayes risk use the full increment, but it is also normalized by the posterior covariance evaluated at $\boldsymbol{x}^k$. When $\boldsymbol{x}^k$ is far from a stationary point for a nonlinear problem, this is not an accurate measure of $\mathbf{P}^a$. The preconditioned increment direction is simply the normalized increment, $\frac{\boldsymbol{v}}{\|\delta \boldsymbol{v}\|}$.

Figure 4.5 shows the Euclidian norm between the direction calculated from the full-rank SVD and those of the same approximation methods as shown in Fig. 4.4. There is a clear benefit to using the LRU across the entire envelope of N_{app} for the exact truncated SVD and all RIOT methods except RIOT-51$^{\top}$. Thus, for nonlinear solvers that use, e.g., DGN or Levenberg-Marquardt type methods, it may be beneficial to apply the LRU for all numbers of modes. As we described previously, the Lanczos recurrence produces identical LRU and LRA increments, as does RIOT-51$^{\top}$. For a given number of approximate modes, $N_{\text{app}} \geq 4$, the increment direction from the Lanczos recurrence is more accurate than any other approach. Once again, the hybridization improves the RSI-enabled RIOT methods to an accuracy better than the truncated SVD.

Although we compared the error behavior of these various Hessian approximation methods, the question of their computational efficiency remains. Figure 4.6 shows the wall-time and CPU time of the three RSI enabled RIOT methods from Fig. 4.4 versus the Euclidian norm of the increment direction error. These plots show the efficiency of each method and approximate mode count to achieve a given error threshold. The times adhere to the assumption that the collective wall-times of the TLM and ADM are approximately$\times 10$ of the NLM, which matches the computational scaling in WRFDA-Chem. Therefore, the number of iterations used in the Lanczos recurrence is found by multiplying its wall-time by 1/10. As indicated by Figs. 4.6(a-c), the increment direction remains the same after approximately 30 Lanczos iterations. Problems that require more iterations to converge will have larger wall-time efficiency improvements than the one tested here.

The hybrid RIOT methods are able to achieve significantly lower error within a given wall-time than the non-hybrid method. Since we are evaluating increment direction, both non-hybrid and hybrid RIOT-51 are more efficient with the LRU across all N_{app} in terms of both wall- and CPU time. If we instead consider the Bayes risk of the full increment (including magnitude), LRA would be more efficient for small N_{app}. RIOT achieves anywhere between a $\times 2$ reduction ($q = 0$, $N_{\text{app}} = 10$) and a $\times 10$ ($q = 0$, $N_{\text{app}} = 60$) reduction in wall-time relative to the Lanczos recurrence. To be clear, that scaling only accounts for time spent in ADM and TLM integrations. As currently implemented in WRFDA-Chem, the DGN line search requires seven sequential NLM

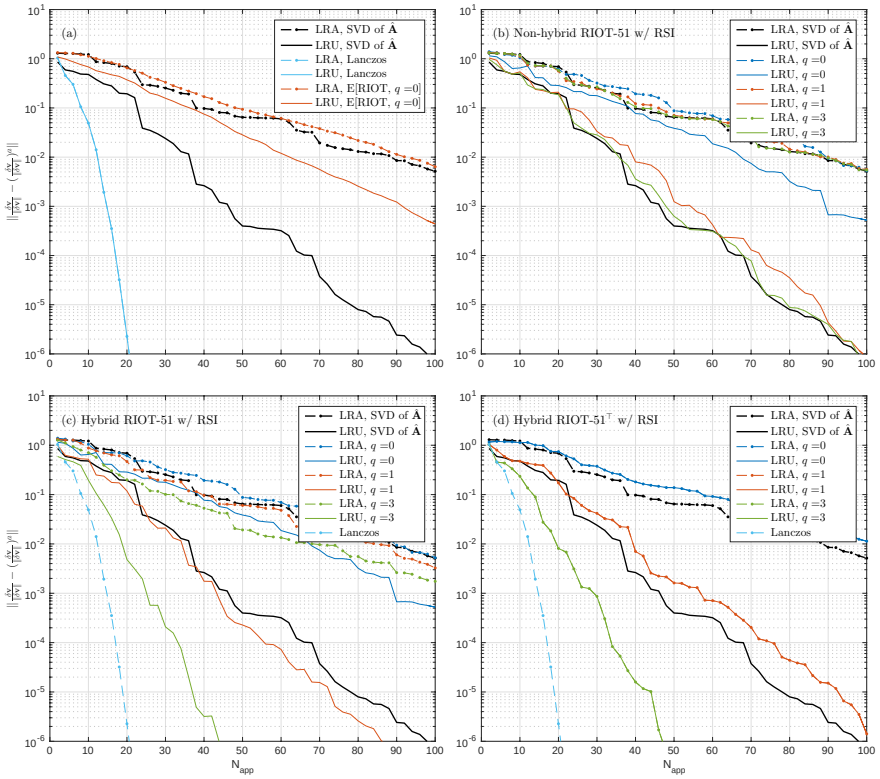

Figure 4.5: Same as Fig. 4.4, except showing the Euclidian norm between the increment direction from using the full-rank SVD of the exact Hessian and those found from several approximation methods.

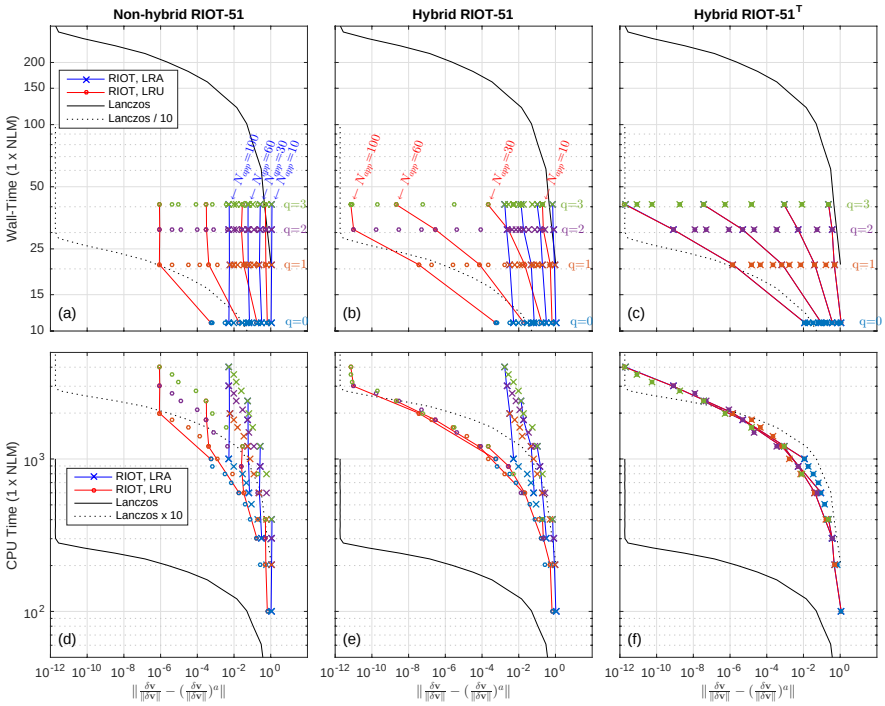

Figure 4.6: Across the rows are wall-time and CPU time of the TLM and ADM portions of a single outer iteration relative to a single simulation of the NLM. All times are plotted versus the Euclidian norm of the increment direction error. Each column is for a different RIOT-based method. Solid red and blue lines connect methods that have the same number of approximation modes, N_{app}. Results are shown for $N_{\text{app}} = [10, 20, \ldots, 100]$. Marker colors refer to different numbers of RSI iterations, q. The Lanczos/10 line shows where $\frac{1}{10}$ of the Lanczos-based wall-time requirement falls. The Lanczos$\times$10 line shows where $\times 10$ of the Lanczos-based CPU time requirement falls.

integrations, which is equivalent to a single call to the ADM. Figures 4.6(a-c) show that causing a wall-time improvement at a specific error threshold requires a minimum number of ensembles greater than $\sim$12 for RIOT-51, while RIOT-51$^\top$ requires only $\sim$8. If computing resources allow, it is always beneficial to increase the number of ensemble members. When the wall-time limits of an application permit only a single RSI iteration ($q = 0$), RIOT-51 performs better than RIOT-51$^\top$. In that instance, the hybrid and non-hybrid implementations are identical. Thus, hybridization is only a benefit when $q > 0$. Figures 4.6(d-e) indicate that all RIOT methods consume at least $\times 5$ as many CPU hours as the Lanczos recurrence in order to reach an equivalent increment direction. When computing resources are limited in terms of CPU hours, and only an increment direction is desired, the choice between RIOT and the Lanczos recurrence is less obvious.

In a perfectly symmetric first outer iteration, the Lanczos recurrence uses the fewest modes to produce an increment at a given error threshold against that produced with the full-rank SVD of the exact Hessian. Still, Bayes risk and direction norms of RIOT increments can be made equivalent by either adding sufficient ensemble members ($\times$2-$\times$5 the number of Lanczos iterations) or utilizing the power iteration and hybridization. While the computational efficiency of higher ensemble counts is associated with higher CPU time, the wall-time gains can be an asset in time-sensitive applications. In practice, the power iteration should only be used when there are not enough processors to run the number of ensembles necessary to achieve a specific performance metric since doing so increases the wall-time of RIOT. If RSI is used, utilizing the hybrid basis has no extra cost, and reduces the number of ensembles needed to converge on an optimal analysis increment. We have not conducted a thorough analysis of how hybridization affects the quality of the basis or the eigenmodes, which are crucial for posterior covariance estimation. In fact, since posterior covariance is evaluated in the last outer iteration of the GN optimization, when $\boldsymbol{b} \approx 0$, using $\hat{\boldsymbol{b}}$ as a basis vector could be detrimental.

The wall-time benefits and CPU costs of posterior covariance estimation is much more balanced between the Lanczos recurrence and RIOT. $\mathbf{P}^a$ typically requires many more sequential Lanczos iterations to converge than the analysis increment. While it is true that methods that

include $\hat{\boldsymbol{b}}$ in the basis require fewer modes than the truncated SVD to converge on optimal analysis increments, the bases of the former remain deficient relative to those of the latter in terms of ϵ_Q. The basis accuracy is especially crucial when calculating posterior variance, which depends solely on the low-rank eigenmodes. Therefore the relatively small values for $\mathrm{LAG_{LZ}}$ in Figs. 4.3(c,f) indicate that RIOT can calculate equivalent posterior covariance with a similar number of modes as the Lanczos recurrence. If the posterior covariance requires 40 to 50 iterations to converge, the wall-time savings from RIOT are 40 to 50 fold, and the additional CPU cost is 5-15% for a standard RIOT-56 or RIOT-51 setup with $q = 0$.

4.3.2 Converged state and posterior covariance

Up to this point we characterized eigenmodes, basis quality, and analysis increments found from the Lanczos recurrence and RIOT for a single outer iteration of incremental 4D-Var. The single outer iteration tests are informative for choosing equivalent numbers of approximate modes between Lanczos and different RIOT configurations. The number of approximate modes is the defining factor in terms of reaching optimality in terms of a convergence. The posterior covariance can be evaluated at that converged state. Since the full emission constraint problem is nonlinear and it is cumbersome to sample enough state vectors to find the global minimum, we do not know the true optimal emissions, nor do we know the true posterior covariance. Instead, we can compare the converged emissions from several scenarios of incremental 4D-Var that use different numbers of inner iterations (Lanczos-GN minimization) or ensemble members (RIOT) in each outer iteration. As a reminder, the focus of this work is to evaluate the equivalence between different inversion methods, and not to provide information about posterior emissions that should be used in future work.

Figure 4.7 shows the posterior BC emissions from wildfires for two emission areas (EA; see Chapter 3 for definitions of specific EAs) where burning sources are significant. All methods are evaluated across 6 outer iterations. As a baseline, the Lanczos recurrence is applied for 30 and 10 inner iterations per outer iteration. RIOT-56 and RIOT-51 (with $q = 0$) are applied with 20,

30, and 60 ensembles. Additionally, we used the Lanczos recurrence, RIOT-56, RIOT-51, and the SVD of the LSR of the Hessian to evaluate posterior variance at the posterior emission state found after 5 outer iterations of Lanczos-GN with $N_{\text{app}} = 30$. The BB emissions start with a factor of $\times 3.8$ uncertainty, which is consistent with the assumptions followed in Chapter 3. The reduction in variance of the emission scaling factors, $\boldsymbol{x}$, compared to the prior is shown in Fig. 4.8. Using the posterior mean and variance together, we can evaluate the various methods.

In both EA2 and EA4, there is a large difference between the Lanczos-GN inversions with either 10 or 30 inner iterations. For EA2, the 10 iteration case exhibits an earlier morning peak, which is inconsistent with the 30 iteration case, as well as most of the RIOT cases. The 10 iteration case also has a slower drop off in afternoon emissions than all other scenarios. Within EA2, RIOT-51 and RIOT-56 match well with the 30 iteration Lanczos-GN case; however, the RIOT-56 case with $N_{\text{app}} = 60$ diverges from the Lanczos-GN 30 iteration case for the early morning peak. The variance reduction for EA2 lines up in time with the largest emission increments. The late morning or early afternoon peak in variance reduction for EA2 is consistent with the fact that the DC-8 measurement platform flew through this region 1 to 2 hours later. In the final outer iteration, the variance reduction of the $N_{\text{app}} = 60$ cases of both RIOT-56 and RIOT-51 agree closely with the LSR SVD in both EA2 and EA4. RIOT-51 with $N_{\text{app}} = 30$ gives equal or higher variance reduction estimates than the Lanczos recurrence, also with $N_{\text{app}} = 30$. Since variance reduction increases monotonically as more modes are used, this could be an indication that RIOT-51 requires as many or fewer modes than the Lanczos recurrence for posterior variance estimation. RIOT-56 may need more approximating modes to converge on equivalent posterior variance to the Lanczos recurrence. Although we enforced symmetry in RIOT-56 by using the SVD of $\mathbf{K}$ in step 16 of Algorithm 3, there may be some impact from applying this algorithm to a Hessian that is not perfectly symmetric. When RIOT-51 and RIOT-56 were applied to the perfectly symmetric Hessian in Sec. 4.3.1, they gave identical results. In EA4, RIOT-56 converges on the same afternoon emission peak found from the Lanczos-GN minimization, but predicts a lower early morning peak by 25%. RIOT-51 exhibits the opposite behavior, exceeding the Lanczos-GN afternoon peak by 9-20%, and converging toward

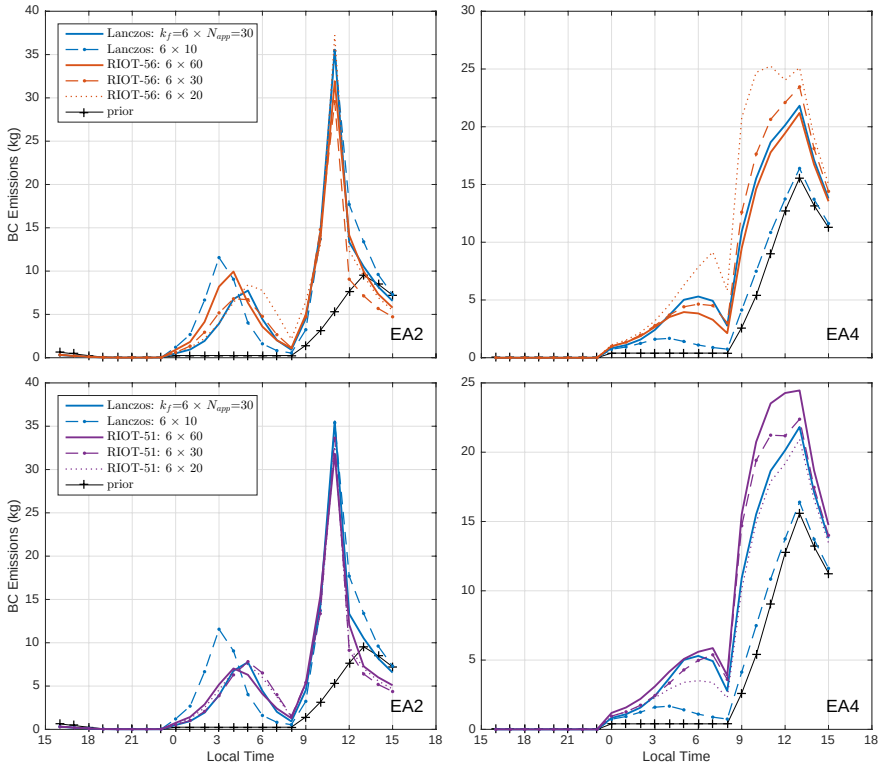

Figure 4.7: Posterior BB emissions of BC for two critical emission areas (EA) using RIOT-56 (top) and RIOT-51 (bottom) compared to the Lanczos recurrence. Each method is applied at multiple numbers of approximate modes (N_{app}) for six outer iterations (k_f).

or slightly exceeding Lanczos-GN in the morning as N_{app} increases. The variance reduction in EA4 indicates that the early morning posterior from the Lanczos-GN minimization is of higher certainty than the afternoon peak. Thus, RIOT-51 is consistent with Lanczos-GN where the observations are most informative.

The posterior anthropogenic BC emissions and scaling factor variance reduction are shown in Figs. 4.9 and 4.10, respectively. The emission area considered (EA6) is centered over Los Angeles, which is the only region where the 22 June flight conducted measurements dedicated to local anthropogenic sources. For the inversion, the anthropogenic emissions start with a factor of $\times 2$ uncertainty, which is consistent with the assumptions followed in Chapter 3. Therefore, the anthropogenic sources have less potential to be adjusted or to have their variance reduced than the BB sources. Additionally, anthropogenic emitters are much more disperse, so that each measurement is much less sensitive to the anthropogenic BC flux in a particular grid cell than it would be to a larger BB flux in the same region. Still, the posterior exhibits large emission increments during the night. With N_{app} =60, both RIOT-56 and RIOT-51 converge close to the Lanczos-GN posterior for anthropogenic emissions. While these increments are potentially influenced by nocturnal boundary layer heights that are difficult to predict with WRF, that does not prevent a comparison between inversion methods. As N_{app} is increased, both RIOT-56 and RIOT-51 converge on the EA6 emissions. As we discussed earlier, the portion of the flight that characterized Los Angeles was primarily in the mid-morning. This explains why EA6 has the largest variance reduction peak at 6:00 LT. RIOT-51 gives equivalent or larger variance reductions than the Lanczos recurrence when both have $N_{\text{app}} = 30$ modes, which is consistent with the results with BB sources. When both use $N_{\text{app}} = 60$ modes, RIOT-51 and RIOT-56 are consistent with the LSR SVD.

Figure 4.6(a-c) showed that the Lanczos recurrence requires 30 or fewer iterations to converge on an increment direction at high precision. Separate tests were conducted with a full Lanczos-GN minimization consisting of six outer iterations with 10 inner iterations each. Although not shown here, the resulting posterior emissions coincided with higher posterior variance than when 30 inner iterations are used. Due to wall-time limitations, we did not conduct Lanczos-based optimizations

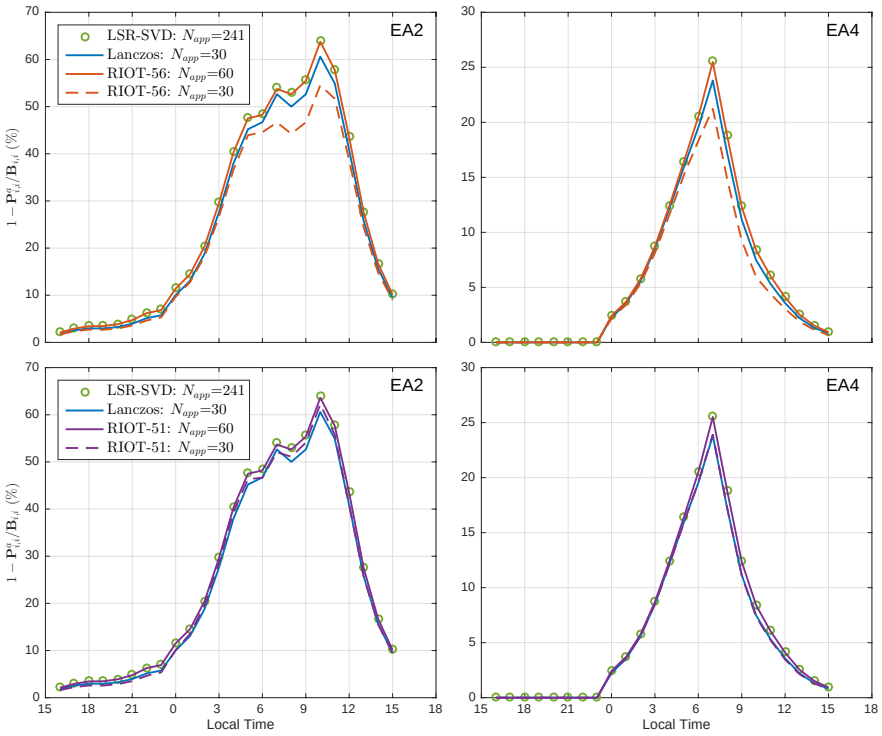

Figure 4.8: Posterior variance reduction for biomass burning scaling factors (%), relative to the prior variance for the same two emission areas as shown in Fig. 4.7.

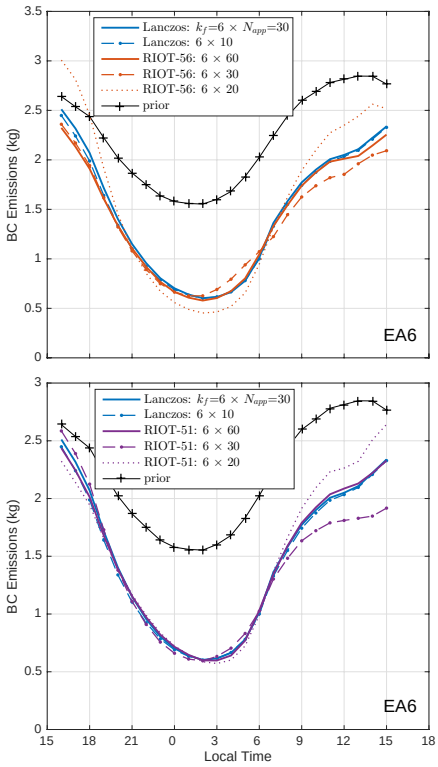

Figure 4.9: Same as Fig. 4.7, but for a single anthropogenic emission area.

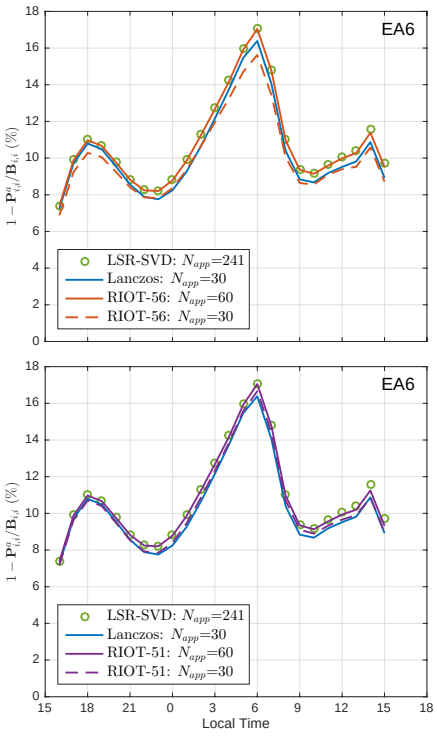

Figure 4.10: Same as Fig. 4.8, but for a single anthropogenic emission area.

for more than 30 iterations.

For the single outer iteration case (see Fig. 4.5), it appeared RIOT (with $q = 0$) required more than $\times 5$ the number of modes as the Lanczos recurrence with $N_{\text{app}} = 10$ to reach increment directions of equal quality. When the Lanczos recurrence uses 30 modes, it appeared RIOT required more than 100 modes with $q = 3$ to reach similar levels of increment direction error. Comparing the final emissions increments on a linear scale yields more promising computational performance for RIOT. The multiple outer iteration tests confirm that many fewer RIOT modes are needed, even when $q = 0$. Contrary to the results shown in Fig. 4.5, 60 ensembles from RIOT and 30 iterations from the Lanczos recurrence produce posteriors that are more consistent with each other than with the posterior produced from 10 Lanczos modes. While these posterior emissions are not identical, their differences could be attributable to errors between the true Hessian of the nonlinear problem and the linearized version used in the optimization. Therefore, when the full nonlinear optimization is considered for this problem, RIOT gives a larger wall-time benefit over Lanczos-GN than would be approximated from a single outer iteration. Furthermore, as was expected from the basis quality produced by each method in Fig. 4.3, the posterior variance consistency between RIOT and Lanczos is promising. Since RIOT with $q = 0$ requires $\leq \times 2$ as many modes as Lanczos for increment and covariance estimation, the CPU cost of RIOT is overblown in Fig. 4.6(d-f). It should be noted that the relative performance for different methods in WRFDA-Chem could be influenced by the initial asymmetry of the Hessian. Although we have added specific algorithmic solutions to enforce symmetry in RIOT, the influence of the initial asymmetry is difficult to assess at this time.

4.4 Conclusions

In this work, we applied a recently proposed parallelized incremental 4D-Var approach, RIOT, to solve for atmospheric chemical emissions. RIOT uses RSVD to approximate the inverse Hessian of the DA cost function, as compared to traditional GN inner loop optimization methods that use sequential iterations. We describe several variations of RIOT; the fastest of these (RIOT-56 and

RIOT-51) only require a single sequential simulation of the adjoint and tangent models per outer loop iteration. We introduce a hybrid stochastic-deterministic power iteration that combines the advantages of RSVD and Krylov subspace methods (e.g., the Lanczos recurrence) by including Hessian basis information contained in the right-hand-side of the least squares problem. RIOT-56 and RIOT-51 were implemented with WRFDA-Chem, which makes the code adaptable to operational weather and air quality forecasting.

We evaluated RIOT and the Lanczos recurrence against the SVD of an exact symmetric Hessian during the first outer iteration of a representative 4D-Var scenario. The eigenspectra of both RSVD and the Lanczos recurrence are greatly improved as more approximation modes are used (i.e., RSVD ensemble members or Lanczos iterations). RSVD is able to produce a basis of equivalent effective rank as the Lanczos recurrence given ten or fewer extra approximate modes. The RSVD basis for a given number of approximate modes improves when the power iteration is used, but so does the wall-time. Bayes risk was used to determine that the Lanczos recurrence requires fewer approximate modes than a truncated SVD of the exact Hessian to converge on an optimal analysis increment. While RIOT-51 and RIOT-56 require many more approximate modes than the truncated SVD, they are also performed in parallel. The hybrid power iteration reduces the mode-dependent error of RIOT to levels similar to those of the Lanczos recurrence.

At a given error threshold in the first outer iteration, the parallel nature of RIOT reduced the wall-time spent in TLM and ADM integrations by $\times 3 - \times 10$ for the problem studied, but also increases CPU time by $\times 5 - \times 10$. We also completed six outer iterations of incremental 4D-Var until convergence was reached at a stationary point. There RIOT produced comparable posterior mean and variance reduction to a Lanczos-GN minimization with as little as $\times 2$ as many modes in regions of high observation information content. Since we could only run the Lanczos-GN algorithm with 30 inner iterations, we do not know for sure that it is at the most optimal local minimum, and it is not necessarily the truth for this case.

At the outset of Sec. 4.3, we declared four sources of error in a GN minimization, which were all discussed except for one: the need to strike a balance between the number of approximate

modes, computational resource limitations, and resolving meaningful information contained in the observations. If too few modes are used, then there is information remaining in the observations that can still be transferred to the posterior. If too many modes are used, then the posterior may contain too much information from the prior (BH16). Here we have characterized the number of approximate modes required by randomized and deterministic in order to converge to an optimal increment, as measured by that which would be given by a full-rank approximation of the Hessian. What still remains is to evaluate exactly what rank of the linearized Hessian is optimal in a nonlinear GN minimization, especially if it is less than rank-n. As discussed by BH16, combining RIOT with preconditioning methods could further reduce the number of approximate modes. Both of these problems emphasize a need to consider the nonlinear problem holistically, with a view beyond independent treatment of each outer iteration.

Chapter 5

Conclusions

The original purpose of this thesis project was to enable 4D-Var in a coupled meteorology and atmospheric chemistry model. The direction of the work evolved as the challenges of reaching that goal presented themselves.

We started in Chapter 2 by developing and applying new tangent linear and adjoint code in WRFPLUS-Chem that treats emissions, dry deposition, PBL mixing, and aging of black carbon (BC) aerosol. These were the minimum pieces necessary to perform 4D-Var inversions for emissions of BC or other nearly inert tracers. Through extensive comparison to finite difference approximations, we were able to debug and successfully implement these linearized models. The most difficult aspect of that work was developing the TLM and ADM. Future work should be focused on adding more linearized model capabilities for treatments of aerosol activation in clouds, wet removal, or aerosol- and gas-phase reaction mechanisms.

One of the practical challenges we foresaw for performing emission inversion was the management of the model trajectory for longer simulations. WRFPLUS and WRFDA were designed to operate in short operational weather forecasting windows of six or twelve hours. However, longer simulations are needed for emission inversions. The existing WRFPLUS code was able to store the model state variables at each time step either in memory or on hard disk. We implemented a second-order checkpointing scheme that balances the I/O and memory requirements of running the TLM and ADM.

With these initial pieces of software developed, we conducted a sensitivity study using the

ADM. First we used WRF-Chem to predict black carbon (BC) aerosol concentrations during the ARCTAS-CARB research campaign in California. We also predicted BC concentrations for an ensemble of model configurations in order to characterize the uncertainty of the model at the times and locations of measurements. We used observations from an SP2 device on a DC-8 aircraft and from filter samples collected at IMPROVE surface sites. These ensemble uncertainties, the instrument uncertainties, and the errors between the observed and predicted concentrations of BC were used to drive the sensitivity tests. The gradients calculated were equivalent to those used in the first stage of 4D-Var, and they are equivalent to the steepest descent direction used in high-dimensional linear optimizations.

In Chapter 3, we modified the incremental 4D-Var framework in WRFDA to conduct inverse modeling of chemical emissions. Incremental 4D-Var is intended to carry out a Gauss-Newton optimization method on fairly linear systems, a limitation heeded in operational weather forecasting. For the most part, the emission and transport mechanisms that affect BC concentrations during dry periods are linear, possibly with the exception of boundary layer mixing. However, the positive-definite nature of BC concentrations and sources means that their errors are lognormally distributed, which required alterations from standard Gaussian assumptions applied in meteorological DA. We treated this nuance with an established exponential emission scaling factor treatment from the literature, which removed the need to use a bounded optimization method, and improved the representativeness of the assumed probability distribution. In doing so, we introduced nonlinearity into an otherwise linear system of equations that caused divergence in the optimization. Then we modified the analysis increment through a damped Gauss-Newton (DGN) line search, which rendered incremental 4D-Var convergent for these unique control variables. We also extended the treatments of control variable covariance from the meteorological variables to the emission scaling factors, which are distributed in space and time. These pieces together combined with the available minimization framework in WRFDA, comprise a new tool, WRFDA-Chem.

WRFDA-Chem was applied during the ARCTAS-CARB period to assess the utility of the surface and aircraft observations in constraining biomass burning and anthropogenic sources of BC

on 22-24 June, 2008. We found that observations on multiple days resulted in temporally and spatially heterogenous emission scaling factors. One reason for this is that the aircraft flew very different flightpaths on the two days, each having a different region of influence in the posterior emissions. Therefore, finding posterior emissions was only the first half of the problem, and we needed to assess the uncertainty reduction relative to the prior that was achievable by using these observations. The posterior variance was calculated using eigenmodes of the cost function Hessian. With these values, we determined that the surface observations had very little information about emissions at the temporal and spatial scales we were trying to constrain, except for in a few model grid cells near Los Angeles and Fresno. We also confirmed that the information content of an observation disperses as it is traced backwards in time through the ADM. In order to perform cross validation of posterior emissions, measurements need to be repeated near the same sources on multiple days, and at multiple times throughout the day.

Due to the wall-time requirements of WRFDA-Chem, we were only able to conduct 4D-Var across 1 to 3 day periods in a single inversion. By exploring this particular period, and with the limitation of a few days of simulation, we suffered a common problem in atmospheric chemical DA, the sparsity of observations. The solution is often to move to longer simulation periods, over which multiple days of wide-coverage satellite measurements can be used. Additionally, through the diagnosis of the trajectory memory constraints in Chapter 2, and subsequent modeling, we determined that simulations with very high resolution or large numbers of chemical species and reactions would have much longer run times that make sequential optimization algorithms intractable. This was a problem that needed to be solved before TLM and ADM treatments of complex chemistry could be worthwhile. All of three of these problems, (1) longer simulation periods, (2) using higher resolution, and (3) treating complex chemical-meteorological interactions, would increase wall-time excessively. In numerical modeling, it is in the interest of the user to push wall-times to their limits, whether those limits are based on monetary costs or turnaround for time-sensitive operational applications. The potential for meaningful results from a simulation increase when we brush those invisible barriers.

In Chapter 4, we aimed at reducing the wall-time of atmospheric 4D-Var inversions to extend the boundary. We applied RIOT in WRFDA-Chem to parallelize the traditionally sequential inner loop of a Gauss Newton optimization procedure. Following recent work in the applied math community, RIOT utilizes independent ensembles of TLM and ADM model evaluations to perform a stochastic approximation of the cost function Hessian. We used a randomized SVD of the Hessian to calculate analysis increments and posterior variance, and compared the results to a sequential Hessian decomposition performed with the Lanczos recurrence. We found wall-time decreases of up to $\times 10$ in the first outer iteration of incremental 4D-Var for equivalent analysis increments. For the full nonlinear optimization consisting of five outer iterations to find the posterior and one to estimate the posterior covariance, the computational scaling improved further. RIOT converged to similar posterior emissions and covariance as the Lanczos-based Gauss Newton optimization. Although this work is ongoing, the implementation of RIOT in WRFDA-Chem is essentially complete. This implementation can be used in research or operationally to reduce 4D-Var wall-times.

Much of this work entailed modifying weather forecasting DA capabilities to suit the needs of atmospheric chemical inverse modeling. At the outset, I expected that all the theory and methods were in place to accomplish the goals of the thesis, and it was my mission to apply them in a new tool. As daunting as that task may have been, there remained a research frontier that needed to be explored in order to enable 4D-Var in a coupled meteorology and atmospheric chemistry model. There are mutual needs across the these two branches of research; this is but one work that attempts to bridge a gap between them.

Bibliography

Jassim Al-Saadi, Amber J. Soja, Robert B. Pierce, James Szykman, Christine Wiedinmyer, Louisa Emmons, Shobha Kondragunta, Xiaoyang Zhang, Chieko Kittaka, Todd Schaack, and Kevin Bowman. Intercomparison of near-real-time biomass burning emissions estimates constrained by satellite fire data. Journal of Applied Remote Sensing, 2(1):021504, 2008. doi: 10.1117/1.2948785.

N. Andela, J. W. Kaiser, G. R. van der Werf, and M. J. Wooster. New fire diurnal cycle characterizations to improve fire radiative energy assessments made from MODIS observations. Atmospheric Chemistry and Physics, 15(15):8831–8846, August 2015. ISSN 1680-7324. doi: 10.5194/acp-15-8831-2015. URL http://www.atmos-chem-phys.net/15/8831/2015/.

Jeffrey L. Anderson. A Local Least Squares Framework for Ensemble Filtering. Mon. Wea. Rev., 131(4):634–642, April 2003. ISSN 0027-0644. doi: 10.1175/1520-0493(2003)131⟨0634:ALLSFF⟩2.0.CO;2. URL http://dx.doi.org/10.1175/1520-0493(2003)131<0634:ALLSFF>2.0.CO;2.

S. C. Anenberg, K. Talgo, S. Arunachalam, P. Dolwick, C. Jang, and J. J. West. Impacts of global, regional, and sectoral black carbon emission reductions on surface air quality and human mortality. Atmospheric Chemistry and Physics, 11(14):7253–7267, July 2011. ISSN 1680-7324. doi: 10.5194/acp-11-7253-2011. URL http://www.atmos-chem-phys.net/11/7253/2011/.

W. M. Angevine, J. Brioude, S. McKeen, and J. S. Holloway. Uncertainty in Lagrangian pollutant transport simulations due to meteorological uncertainty from a mesoscale WRF ensemble. Geoscientific Model Development, 7(6):2817–2829, December 2014. ISSN 1991-9603. doi: 10.5194/gmd-7-2817-2014. URL http://www.geosci-model-dev.net/7/2817/2014/.

Thomas Auligné, Benjamin Ménétrier, Andrew C. Lorenc, and Mark Buehner. Ensemble-Variational Integrated Localized Data Assimilation. Monthly Weather Review, June 2016. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-15-0252.1. URL http://journals.ametsoc.org/doi/10.1175/MWR-D-15-0252.1.

A. Baklanov, K. . Schlünzen, P. Suppan, J. Baldasano, D. Brunner, S. Aksoyoglu, G. Carmichael, J. Douros, J. Flemming, R. Forkel, S. Galmarini, M. Gauss, G. Grell, M. Hirtl, S. Joffre, O. Jorba, E. Kaas, M. Kaasik, G. Kallos, X. Kong, U. Korsholm, A. Kurganskiy, J. Kushta, U. Lohmann, A. Mahura, A. Manders-Groot, A. Maurizi, N. Moussiopoulos, S. T. Rao, N. Savage, C. Seigneur, R. S. Sokhi, E. Solazzo, S. Solomos, B. Sørensen, G. Tsegas, E. Vignati, B. Vogel, and Y. Zhang. Online coupled regional meteorology chemistry models in Europe: current status and prospects. Atmospheric Chemistry and Physics, 14(1):317–398, January 2014. ISSN 1680-7324. doi: 10.5194/acp-14-317-2014. URL http://www.atmos-chem-phys.net/14/317/2014/.

Dale Barker, M.-S. Lee, Y.-R. Guo, W. Huang, H. Huang, and Q. Rizvi. WRF-Var - A unified 3/4d-Var variational data assimilation system for WRF. In Sixth WRF/15th MM5 Users' Workshop, page 17, Boulder, CO, NCAR, June 2005. URL http://www2.mmm.ucar.edu/wrf/users/workshops/WS2005/presentations/session10/1-Barker.pdf.

Dale M. Barker, W. Huang, Yong-Run Guo, A. J. Bourgeois, and Q. N. Xiao. A three-dimensional variational data assimilation system for MM5: Implementation and initial results. Monthly Weather Review, 132(4):897–914, 2004. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(2004)132%3C0897:ATVDAS%3E2.0.CO;2.

Stanley L. Barnes. A Technique for Maximizing Details in Numerical Weather Map Analysis. J. Appl. Meteor., 3(4):396–409, August 1964. ISSN 0021-8952. doi: 10.1175/1520-0450(1964)003⟨0396:ATFMDI⟩2.0.CO;2. URL http://dx.doi.org/10.1175/1520-0450(1964)003<0396:ATFMDI>2.0.CO;2.

A. Benedetti, J.-J. Morcrette, O. Boucher, A. Dethof, R. J. Engelen, M. Fisher, H. Flentje, N. Huneeus, L. Jones, J. W. Kaiser, S. Kinne, A. Mangold, M. Razinger, A. J. Simmons, and M. Suttie. Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: 2. Data assimilation. Journal of Geophysical Research, 114(D13205):1–18, July 2009. ISSN 0148-0227. doi: 10.1029/2008JD011115. URL http://doi.wiley.com/10.1029/2008JD011115.

Peter Bergamaschi, Christian Frankenberg, Jan Fokke Meirink, Maarten Krol, M. Gabriella Villani, Sander Houweling, Frank Dentener, Edward J. Dlugokencky, John B. Miller, Luciana V. Gatti, Andreas Engel, and Ingeborg Levin. Inverse modeling of global and regional CH4 emissions using SCIAMACHY satellite retrievals. Journal of Geophysical Research, 114(D22301):1–28, November 2009. ISSN 0148-0227. doi: 10.1029/2009JD012287. URL http://doi.wiley.com/10.1029/2009JD012287.

Å. Björck. Numerics of Gram-Schmidt orthogonalization. Linear Algebra and its Applications, 197:297–316, January 1994. ISSN 0024-3795. doi: 10.1016/0024-3795(94)90493-6. URL http://www.sciencedirect.com/science/article/pii/0024379594904936.

T. C. Bond, S. J. Doherty, D. W. Fahey, P. M. Forster, T. Berntsen, B. J. DeAngelo, M. G. Flanner, S. Ghan, B. Kärcher, D. Koch, S. Kinne, Y. Kondo, P. K. Quinn, M. C. Sarofim, M. G. Schultz, M. Schulz, C. Venkataraman, H. Zhang, S. Zhang, N. Bellouin, S. K. Guttikunda, P. K. Hopke, M. Z. Jacobson, J. W. Kaiser, Z. Klimont, U. Lohmann, J. P. Schwarz, D. Shindell, T. Storelvmo, S. G. Warren, and C. S. Zender. Bounding the role of black carbon in the climate system: A scientific assessment: BLACK CARBON IN THE CLIMATE SYSTEM. Journal of Geophysical Research: Atmospheres, 118(11):5380–5552, June 2013. ISSN 2169897X. doi: 10.1002/jgrd.50171. URL http://doi.wiley.com/10.1002/jgrd.50171.

L. Boschetti, H. D. Eva, P. A. Brivio, and J. M. Grégoire. Lessons to be learned from the comparison of three satellite-derived biomass burning products. Geophysical Research Letters, 31(21):n/a–n/a, 2004. ISSN 1944-8007. doi: 10.1029/2004GL021229. URL http://dx.doi.org/10.1029/2004GL021229.

Nicolas Bousserez and Daven K. Henze. Optimal and scalable methods to approximate the solutions of large-scale Bayesian problems: Theory and application to atmospheric inversions and data assimilation. arXiv preprint arXiv:1609.06431, 2016. URL http://arxiv.org/abs/1609.06431.

Jerome Brioude, W. M. Angevine, Ravan Ahmadov, S-W Kim, Stephanie Evan, S. A. McKeen, E-Y Hsie, G. J. Frost, J. A. Neuman, and I. B. Pollack. Top-down estimate of surface flux in the Los Angeles Basin using a mesoscale inverse modeling technique: assessing anthropogenic emissions of CO, NO x and CO 2 and their impacts. Atmospheric Chemistry and Physics Discussions, 12 (12):31439–31481, 2012. URL http://www.atmos-chem-phys-discuss.net/12/31439/2012/.

K. Brown, I. Gejadze, and A. Ramage. A Multilevel Approach for Computing the Limited-Memory Hessian and its Inverse in Variational Data Assimilation. SIAM J. Sci. Comput., pages A2934–A2963, January 2016. ISSN 1064-8275. doi: 10.1137/15M1041407. URL http://dx.doi.org/10.1137/15M1041407.

Mark Buehner. Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. Quarterly Journal of the Royal Meteorological Society, 131(607):1013–1043, 2005. ISSN 1477-870X. doi: 10.1256/qj.04.15. URL http://dx.doi.org/10.1256/qj.04.15.

R. Byrd, G. Chin, W. Neveitt, and J. Nocedal. On the Use of Stochastic Hessian Information in Optimization Methods for Machine Learning. SIAM J. Optim., 21(3):977–995, July 2011. ISSN 1052-6234. doi: 10.1137/10079923X. URL http://dx.doi.org/10.1137/10079923X.

Gregory R Carmichael, Adrian Sandu, Tianfeng Chai, Dacian N Daescu, Emil M Constantinescu, and Youhua Tang. Predicting air quality: Improvements through advanced methods to integrate models and measurements. Journal of Computational Physics, 227(7):3540–3571, 2008. URL http://www.sciencedirect.com/science/article/pii/S0021999107000836.

J. Charney. The Use of the Primitive Equations of Motion in Numerical Prediction. Tellus, 7(1): 22–26, 1955. ISSN 2153-3490. doi: 10.1111/j.2153-3490.1955.tb01138.x. URL http://dx.doi.org/10.1111/j.2153-3490.1955.tb01138.x.

J. G. Charney, R. Fjörtoft, and J. Von Neumann. Numerical Integration of the Barotropic Vorticity Equation. Tellus, 2(4):237–254, 1950. ISSN 2153-3490. doi: 10.1111/j.2153-3490.1950.tb00336.x. URL http://dx.doi.org/10.1111/j.2153-3490.1950.tb00336.x.

Yan Chen and Dean S. Oliver. Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. Computational Geosciences, 17 (4):689–703, August 2013. ISSN 1420-0597, 1573-1499. doi: 10.1007/s10596-013-9351-5. URL http://link.springer.com/10.1007/s10596-013-9351-5.

Mian Chin, Dennis L. Savoie, Barry J. Huebert, Alan R. Bandy, Donald C. Thornton, Timothy S. Bates, Patricia K. Quinn, Eric S. Saltzman, and Warren J. De Bruyn. Atmospheric sulfur cycle simulated in the global model GOCART: Comparison with field observations and regional budgets. Journal of Geophysical Research: Atmospheres, 105(D20):24689–24712, 2000. ISSN 2156-2202. doi: 10.1029/2000JD900385. URL http://dx.doi.org/10.1029/2000JD900385.

Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In Proceedings of the forty-first annual ACM symposium on Theory of computing, pages 205–214, Bethesda, MD, USA, 2009. ACM. ISBN 978-1-60558-506-2.

A. M. Clayton, A. C. Lorenc, and D. M. Barker. Operational implementation of a hybrid ensemble/4d-Var global data assimilation system at the Met Office. Quarterly Journal of the

Royal Meteorological Society, 139(675):1445–1461, 2013. ISSN 1477-870X. doi: 10.1002/qj.2054. URL http://dx.doi.org/10.1002/qj.2054.

Great Britain Coal Commission and G.D.C. Argyll. Report of the Commissioners appointed to inquire into the several matters relating to coal in the United Kingdom. Number v. 3 in C (Series) (Great Britain. Parliament). Printed by G.E. Eyre and W. Spottiswoode for H.M. Stationery off., 1871. URL https://books.google.com/books?id=epMNAAAAYAAJ.

P. Courtier, J.-N. Thépaut, and A. Hollingsworth. A strategy for operational implementation of 4d-Var, using an incremental approach. Quarterly Journal of the Royal Meteorological Society, 120(519):1367–1387, 1994. ISSN 1477-870X. doi: 10.1002/qj.49712051912. URL http://dx.doi.org/10.1002/qj.49712051912.

Yu Yan Cui, Jerome Brioude, Stuart A. McKeen, Wayne M. Angevine, Si-Wan Kim, Gregory J. Frost, Ravan Ahmadov, Jeff Peischl, Nicolas Bousserez, Zhen Liu, Thomas B. Ryerson, Steve C. Wofsy, Gregory W. Santoni, Eric A. Kort, Marc L. Fischer, and Michael Trainer. Top-down estimate of methane emissions in California using a mesoscale inverse modeling technique: The South Coast Air Basin: Inverse Estimate of Methane Emissions. Journal of Geophysical Research: Atmospheres, 120(13):6698–6711, July 2015. ISSN 2169897X. doi: 10.1002/2014JD023002. URL http://doi.wiley.com/10.1002/2014JD023002.

Anton S. Darmenov and Arlindo da Silva. The Quick Fire Emissions Dataset (QFED)–Documentation of versions 2.1, 2.2, and 2.4. Technical Report NASA TM-2013-104606/Vol 32, NASA, April 2013. URL http://gmao.gsfc.nasa.gov/pubs/tm/.

Gérald Desroziers and Loïk Berre. Accelerating and parallelizing minimizations in ensemble and deterministic variational assimilations. Quarterly Journal of the Royal Meteorological Society, 138(667):1599–1610, 2012. ISSN 1477-870X. doi: 10.1002/qj.1886. URL http://dx.doi.org/10.1002/qj.1886.

François-xavier le Dimet and Olivier Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus A, 38A(2):97–110, 1986. ISSN 1600-0870. doi: 10.1111/j.1600-0870.1986.tb00459.x. URL http://dx.doi.org/10.1111/j.1600-0870.1986.tb00459.x.

D.J. Diner, J.C. Beckert, T.H. Reilly, C.J. Bruegge, J.E. Conel, R.A. Kahn, J.V. Martonchik, T.P. Ackerman, R. Davies, S.A.W. Gerstl, H.R. Gordon, J.-P. Muller, R.B. Myneni, P.J. Sellers, B. Pinty, and M.M. Verstraete. Multi-angle Imaging SpectroRadiometer (MISR) instrument description and experiment overview. Geoscience and Remote Sensing, IEEE Transactions on, 36(4):1072–1087, July 1998. ISSN 0196-2892. doi: 10.1109/36.700992.

O. Dubovik, T. Lapyonok, Y. J. Kaufman, M. Chin, P. Ginoux, R. A. Kahn, and A. Sinyuk. Retrieving global aerosol sources from satellites using inverse modeling. Atmospheric Chemistry and Physics, 8(2):209–250, 2008. URL http://www.atmos-chem-phys.net/8/209/2008/acp-8-209-2008.html.

H. Elbern, A. Strunk, H. Schmidt, and O. Talagrand. Emission rate and chemical state estimation by 4-dimensional variational inversion. Atmospheric Chemistry and Physics, 7(14):3749–3769, 2007. URL http://www.atmos-chem-phys.net/7/3749/.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans, 99(C5): 10143–10162, 1994. ISSN 2156-2202. doi: 10.1029/94JC00572. URL http://dx.doi.org/10.1029/94JC00572.

Geir Evensen and Peter Jan Van Leeuwen. An ensemble Kalman smoother for nonlinear dynamics. Monthly Weather Review, 128(6):1852–1867, 2000. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(2000)128%3C1852%3AAEKSFN%3E2.0.CO%3B2.

M. Fisher and P. Courtier. Estimating the covariance matrices of analysis and forecast error in variational data assimilation. Technical Report Technical Memorandum 220, European Centre for Medium-Range Weather Forecasts, August 1995.

M. Fisher, S. Gratton, S. Gürol, Y. Trémolet, and X. Vasseur. Low rank updates in preconditioning the saddle point systems arising from data assimilation problems. Technical Report TR/PA/16/135, Cerfacs, May 2016.

Mike Fisher, Yannick Tremolet, Harri Auvinen, David Tan, and Paul Poli. Weak-constraint and long-window 4d-Var. Techn. Rep, 655, 2011. URL http://www.ecmwf.int/sites/default/files/elibrary/2011/9414-weak-constraint-and-long-window-4dvar.pdf.

J. Flemming, A. Inness, H. Flentje, V. Huijnen, P. Moinat, M. G. Schultz, and O. Stein. Coupling global chemistry transport models to ECMWF's integrated forecast system. Geoscientific Model Development, 2(2):253–265, 2009. URL http://www.geosci-model-dev.net/2/253/.

Steven J. Fletcher and Andrew S. Jones. Multiplicative and Additive Incremental Variational Data Assimilation for Mixed Lognormal–Gaussian Errors. Monthly Weather Review, 142(7): 2521–2544, July 2014. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-13-00136.1. URL http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-13-00136.1.

Steven J. Fletcher and Milija Zupanski. Implications and impacts of transforming lognormal variables into normal variables in VAR. Meteorologische Zeitschrift, 16(6):755–765, 2007. URL http://www.ingentaconnect.com/content/schweiz/mz/2007/00000016/00000006/art00016.

S. R. Freitas, K. M. Longo, R. Chatfield, D. Latham, M. A. F. Silva Dias, M. O. Andreae, E. Prins, J. C. Santos, R. Gielow, and J. A. Carvalho Jr. Including the sub-grid scale plume rise of vegetation fires in low resolution atmospheric transport models. Atmospheric Chemistry and Physics, 7(13):3385–3398, 2007. URL http://www.atmos-chem-phys.net/7/3385/2007/acp-7-3385-2007.html.

S. R. Freitas, K. M. Longo, J. Trentmann, and D. Latham. Technical Note: Sensitivity of 1-D smoke plume rise models to the inclusion of environmental wind drag. Atmospheric Chemistry and Physics, 10(2):585–594, 2010. URL http://www.atmos-chem-phys.net/10/585/2010/acp-10-585-2010.pdf.

S. R. Freitas, K. M. Longo, M. F. Alonso, M. Pirre, V. Marecal, G. Grell, R. Stockler, R. F. Mello, and M. Sánchez Gácita. PREP-CHEM-SRC – 1.0: a preprocessor of trace gas and aerosol emission fields for regional and global atmospheric chemistry models. Geoscientific Model Development, 4(2):419–433, May 2011. ISSN 1991-9603. doi: 10.5194/gmd-4-419-2011. URL http://www.geosci-model-dev.net/4/419/2011/.

J. S. Fu, N. C. Hsu, Y. Gao, K. Huang, C. Li, N.-H. Lin, and S.-C. Tsay. Evaluating the influences of biomass burning during 2006 BASE-ASIA: a regional chemical transport modeling. Atmospheric Chemistry and Physics, 12(9):3837–3855, May 2012. ISSN 1680-7324. doi: 10.5194/acp-12-3837-2012. URL http://www.atmos-chem-phys.net/12/3837/2012/.

L.S. Gandin. Objective analysis of meteorological fields. Translated from the Russian. Jerusalem (Israel Program for Scientific Translations), 1965. Quarterly Journal of the Royal Meteorological Society, 92(393):447–447, 1966. ISSN 1477-870X. doi: 10.1002/qj.49709239320. URL http://dx.doi.org/10.1002/qj.49709239320.

M. Ghil, S. Cohn, J. Tavantzis, K. Bube, and E. Isaacson. Applications of Estimation Theory to Numerical Weather Prediction. In Lennart Bengtsson, Michael Ghil, and Erland K?§ll?©n, editors, Dynamic Meteorology: Data Assimilation Methods, volume 36 of Applied Mathematical Sciences, pages 139–224. Springer New York, 1981. ISBN 978-0-387-90632-4. URL http://dx.doi.org/10.1007/978-1-4612-5970-1_5.

Ralf Giering and Thomas Kaminski. Recipes for Adjoint Code Construction. ACM Trans. Math. Softw., 24(4):437–474, December 1998. ISSN 0098-3500. doi: 10.1145/293686.293695. URL http://doi.acm.org/10.1145/293686.293695.

B. Gilchrist and G. P. Cressman. An Experiment in Objective Analysis. Tellus, 6(4):309–318, 1954. ISSN 2153-3490. doi: 10.1111/j.2153-3490.1954.tb01126.x. URL http://dx.doi.org/10.1111/j.2153-3490.1954.tb01126.x.

G.H. Golub and C.F. Van Loan. Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1996. ISBN 978-0-8018-5414-9.

Thomas J. Grahame, Rebecca Klemm, and Richard B. Schlesinger. Public health and components of particulate matter: The changing assessment of black carbon. Journal of the Air & Waste Management Association, 64(6):620–660, 2014. doi: 10.1080/10962247.2014.912692. URL http://dx.doi.org/10.1080/10962247.2014.912692.

S. Gratton, A. S. Lawless, and N. K. Nichols. Approximate Gauss–Newton Methods for Nonlinear Least Squares Problems. SIAM Journal on Optimization, 18(1):106–132, January 2007. ISSN 1052-6234, 1095-7189. doi: 10.1137/050624935. URL http://epubs.siam.org/doi/abs/10.1137/050624935.

Serge Gratton, Selime Gürol, and Philippe L. Toint. Preconditioning and globalizing conjugate gradients in dual space for quadratically penalized nonlinear-least squares problems. Computational Optimization and Applications, 54(1):1–25, January 2013. ISSN 0926-6003, 1573-2894. doi: 10.1007/s10589-012-9478-7. URL http://link.springer.com/10.1007/s10589-012-9478-7.

G. Grell, S. R. Freitas, M. Stuefer, and J. Fast. Inclusion of biomass burning in WRF-Chem: impact of wildfires on weather forecasts. Atmospheric Chemistry and Physics, 11(11):5289–5303, June 2011. ISSN 1680-7324. doi: 10.5194/acp-11-5289-2011. URL http://www.atmos-chem-phys.net/11/5289/2011/.

G. A. Grell and S. R. Freitas. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. Atmospheric Chemistry and Physics, 14(10):5233–5250, May 2014. ISSN 1680-7324. doi: 10.5194/acp-14-5233-2014. URL http://www.atmos-chem-phys.net/14/5233/2014/.

Georg Grell and Alexander Baklanov. Integrated modeling for forecasting weather and air quality: A call for fully coupled approaches. Atmospheric Environment, 45(38):6845 – 6851, 2011. ISSN 1352-2310. doi: http://dx.doi.org/10.1016/j.atmosenv.2011.01.017. URL http://www.sciencedirect.com/science/article/pii/S1352231011000240. Modeling of Air Quality Impacts, Forecasting and Interactions with Climate.

Georg A. Grell, Richard Knoche, Steven E. Peckham, and Stuart A. McKeen. Online versus offline air quality modeling on cloud-resolving scales. Geophysical Research Letters, 31(L16117): 1–4, 2004. ISSN 1944-8007. doi: 10.1029/2004GL020175. URL http://dx.doi.org/10.1029/2004GL020175.

Georg A Grell, Steven E Peckham, Rainer Schmitz, Stuart A McKeen, Gregory Frost, William C Skamarock, and Brian Eder. Fully coupled "online" chemistry within the WRF model. Atmospheric Environment, 39(37):6957–6975, 2005. URL http://www.sciencedirect.com/science/article/pii/S1352231005003560.

Andrew P. Grieshop, Eric M. Lipsky, Natalie J. Pekney, Satoshi Takahama, and Allen L. Robinson. Fine particle emission factors from vehicles in a highway tunnel: Effects of fleet composition and season. Atmospheric Environment, 40, Supplement 2:287 – 298, 2006. ISSN 1352-2310. doi: http://dx.doi.org/10.1016/j.atmosenv.2006.03.064. URL http://www.sciencedirect.com/science/article/pii/S1352231006005796.

J. J. Guerrette and D. K. Henze. Development and application of the WRFPLUS-Chem online chemistry adjoint and WRFDA-Chem assimilation system. Geoscientific Model Development, 8(6):1857–1876, 2015. doi: 10.5194/gmd-8-1857-2015. URL http://www.geosci-model-dev.net/8/1857/2015/.

Liam Gumley. MODIS Today. Space Science and Engineering Center, University of Wisconsin-Madison, Madison, WI, USA, June 2008. URL http://ge.ssec.wisc.edu/modis-today.

S. Gürol, A. T. Weaver, A. M. Moore, A. Piacentini, H. G. Arango, and S. Gratton. B-preconditioned minimization algorithms for variational data assimilation with the dual formulation. Quarterly Journal of the Royal Meteorological Society, 140(679):539–556, 2014. ISSN 1477-870X. doi: 10.1002/qj.2150. URL http://dx.doi.org/10.1002/qj.2150.

A. Hakami, D. K. Henze, J. H. Seinfeld, T. Chai, Y. Tang, G. R. Carmichael, and A. Sandu. Adjoint inverse modeling of black carbon during the Asian Pacific Regional Aerosol Characterization Experiment. Journal of Geophysical Research: Atmospheres, 110(D14301):1–17, 2005. ISSN 2156-2202. doi: 10.1029/2004JD005671. URL http://dx.doi.org/10.1029/2004JD005671.

N. Halko, P. G. Martinsson, and J. A. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. SIAM Review, 53(2):217–288, January 2011. ISSN 0036-1445, 1095-7200. doi: 10.1137/090771806. URL http://epubs.siam.org/doi/abs/10.1137/090771806.

Leigh J Halliwell. The Lognormal Random Multivariate. In Casualty Actuarial Society E-Forum, Spring 2015, page 5, Arlington, Virginia, 2015. URL http://www.casact.org/pubs/forum/15spforum/Halliwell.pdf.

Thomas M. Hamill and Chris Snyder. A Hybrid Ensemble Kalman Filter–3d Variational Analysis Scheme. Mon. Wea. Rev., 128(8):2905–2919, August 2000. ISSN 0027-0644. doi:

10.1175/1520-0493(2000)128⟨2905:AHEKFV⟩2.0.CO;2. URL http://dx.doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2.

J. Hansen, M. Sato, and R. Ruedy. Radiative forcing and climate response. Journal of Geophysical Research: Atmospheres, 102(D6):6831–6864, 1997. ISSN 2156-2202. doi: 10.1029/96JD03436. URL http://dx.doi.org/10.1029/96JD03436.

M. Hansen, R. DeFries, J.R. Townshend, M. Carroll, C. Dimiceli, and R. Sohlberg. 500 m MODIS Vegetation Continuous Fields. The Global Land Cover Facility, College Park, Maryland, 2003.

Laurent Hascoet and Valérie Pascual. The Tapenade Automatic Differentiation Tool: Principles, Model, and Specification. ACM Trans. Math. Softw., 39(3):20:1–20:43, May 2013. ISSN 0098-3500. doi: 10.1145/2450153.2450158. URL http://doi.acm.org/10.1145/2450153.2450158.

D. K. Henze, A. Hakami, J. H. Seinfeld, and others. Development of the adjoint of GEOS-Chem. Atmospheric Chemistry and Physics, 7(9):2413–2433, 2007. URL http://hal.archives-ouvertes.fr/hal-00296220/.

D. K. Henze, J. H. Seinfeld, and D. T. Shindell. Inverse modeling and mapping US air quality influences of inorganic $PM_{2.5}$ precursor emissions using the adjoint of GEOS-Chem. Atmospheric Chemistry and Physics, 9(16):5877–5903, 2009. doi: 10.5194/acp-9-5877-2009. URL http://www.atmos-chem-phys.net/9/5877/2009/.

P. L. Houtekamer, Louis Lefaivre, Jacques Derome, Harold Ritchie, and Herschel L. Mitchell. A System Simulation Approach to Ensemble Prediction. Mon. Wea. Rev., 124(6):1225–1242, June 1996. ISSN 0027-0644. doi: 10.1175/1520-0493(1996)124⟨1225:ASSATE⟩2.0.CO;2. URL http://dx.doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

Cho-Jui Hsieh and Peder A. Olsen. Nuclear Norm Minimization via Active Subspace Selection. In 31st International Conference on Machine Learning, volume 32, pages 575–583, Beijing, China, 2014. JMLR. URL http://www.jmlr.org/proceedings/papers/v32/hsiehb14.pdf.

Xiang-Yu Huang, Qingnong Xiao, Dale M. Barker, Xin Zhang, John Michalakes, Wei Huang, Tom Henderson, John Bray, Yongsheng Chen, Zaizhong Ma, Jimy Dudhia, Yongrun Guo, Xiaoyan Zhang, Duk-Jin Won, Hui-Chuan Lin, and Ying-Hwa Kuo. Four-Dimensional Variational Data Assimilation for WRF: Formulation and Preliminary Results. Monthly Weather Review, 137(1):299–314, January 2009. ISSN 0027-0644. doi: 10.1175/2008MWR2577.1. URL http://dx.doi.org/10.1175/2008MWR2577.1.

N. Huneeus, O. Boucher, and F. Chevallier. Simplified aerosol modeling for variational data assimilation. Geoscientific Model Development, 2(2):213–229, 2009. URL http://www.geosci-model-dev.net/2/213/2009/gmd-2-213-2009.html.

Charles Ichoku, Ralph Kahn, and Mian Chin. Satellite contributions to the quantitative characterization of biomass burning for climate modeling. Atmospheric Research, 111:1 – 28, 2012. ISSN 0169-8095. doi: http://dx.doi.org/10.1016/j.atmosres.2012.03.007. URL http://www.sciencedirect.com/science/article/pii/S0169809512000750.

A. Inness, F. Baier, A. Benedetti, I. Bouarar, S. Chabrillat, H. Clark, C. Clerbaux, P. Coheur, R. J. Engelen, Q. Errera, J. Flemming, M. George, C. Granier, J. Hadji-Lazaro, V. Huijnen, D. Hurtmans, L. Jones, J. W. Kaiser, J. Kapsomenakis, K. Lefever, J. Leitão, M. Razinger,

A. Richter, M. G. Schultz, A. J. Simmons, M. Suttie, O. Stein, J.-N. Thépaut, V. Thouret, M. Vrekoussis, C. Zerefos, and the MACC team. The MACC reanalysis: an 8 yr data set of atmospheric composition. Atmospheric Chemistry and Physics, 13(8):4073–4109, April 2013. ISSN 1680-7324. doi: 10.5194/acp-13-4073-2013. URL http://www.atmos-chem-phys.net/13/4073/2013/.

D. J. Jacob, J. H. Crawford, H. Maring, A. D. Clarke, J. E. Dibb, L. K. Emmons, R. A. Ferrare, C. A. Hostetler, P. B. Russell, H. B. Singh, A. M. Thompson, G. E. Shaw, E. McCauley, J. R. Pederson, and J. A. Fisher. The Arctic Research of the Composition of the Troposphere from Aircraft and Satellites (ARCTAS) mission: design, execution, and first results. Atmospheric Chemistry and Physics, 10(11):5191–5212, 2010. doi: 10.5194/acp-10-5191-2010. URL http://www.atmos-chem-phys.net/10/5191/2010/.

Nicole AH Janssen, Mariam E Gerlofs-Nijland, Timo Lanki, Raimo O Salonen, Flemming Cassee, Gerard Hoek, Paul Fischer, Bert Brunekreef, and Michal Krzyzanowski. Health effects of black carbon. World Health Organization, Regional Office for Europe, Copenhagen, 2012. ISBN 978 92 890 0265 3. URL http://www.euro.who.int/__data/assets/pdf_file/0004/162535/e96541.pdf. OCLC: 930804705.

Z. Jiang, D. B. A. Jones, H. M. Worden, and D. K. Henze. Sensitivity of top-down CO source estimates to the modeled vertical structure in atmospheric CO. Atmospheric Chemistry and Physics, 15(3):1521–1537, February 2015. ISSN 1680-7324. doi: 10.5194/acp-15-1521-2015. URL http://www.atmos-chem-phys.net/15/1521/2015/.

W. Matt Jolly, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, and David M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. Nat Commun, 6, July 2015. URL http://dx.doi.org/10.1038/ncomms8537.

J. W. Kaiser, A. Heil, M. O. Andreae, A. Benedetti, N. Chubarova, L. Jones, J.-J. Morcrette, M. Razinger, M. G. Schultz, M. Suttie, and G. R. van der Werf. Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power. Biogeosciences, 9(1):527–554, 2012. doi: 10.5194/bg-9-527-2012. URL http://www.biogeosciences.net/9/527/2012/.

E. Kalnay. Atmospheric Modeling, Data Assimilation, and Predictability. Cambridge University Press, 2003. ISBN 978-0-521-79179-3. URL https://books.google.com/books?id=zx_BakP2I5gC.

J. W. Kaminski, L. Neary, J. Struzewska, J. C. McConnell, A. Lupu, J. Jarosz, K. Toyota, S. L. Gong, J. Côté, X. Liu, K. Chance, and A. Richter. GEM-AQ, an on-line global multiscale chemical weather modelling system: model description and evaluation of gas phase chemistry processes. Atmospheric Chemistry and Physics, 8(12):3255–3281, 2008. doi: 10.5194/acp-8-3255-2008. URL http://www.atmos-chem-phys.net/8/3255/2008/.

C.T. Kelley. Iterative Methods for Optimization. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, 1999. ISBN 978-0-89871-433-3.

S.-W. Kim, B. C. McDonald, S. Baidar, S. S. Brown, B. Dube, R. A. Ferrare, G. J. Frost, R. A. Harley, J. S. Holloway, H.-J. Lee, S. A. McKeen, J. A. Neuman, J. B. Nowak, H. Oetjen, I. Ortega,

I. B. Pollack, J. M. Roberts, T. B. Ryerson, A. J. Scarino, C. J. Senff, R. Thalman, M. Trainer, R. Volkamer, N. Wagner, R. A. Washenfelder, E. Waxman, and C. J. Young. Modeling the weekly cycle of NOx and CO emissions and their impacts on O3 in the Los Angeles-South Coast Air Basin during the CalNex 2010 field campaign: Modeling Weekly Cycle of the LA Air Quality in 2010. Journal of Geophysical Research: Atmospheres, 121(3):1340–1360, February 2016. ISSN 2169897X. doi: 10.1002/2015JD024292. URL http://doi.wiley.com/10.1002/2015JD024292.

M.D. King, Y.J. Kaufman, W.P. Menzel, and D. Tanre. Remote sensing of cloud, aerosol, and water vapor properties from the moderate resolution imaging spectrometer (MODIS). Geoscience and Remote Sensing, IEEE Transactions on, 30(1):2–27, January 1992. ISSN 0196-2892. doi: 10.1109/36.124212.

D. E. Kinnison, G. P. Brasseur, S. Walters, R. R. Garcia, D. R. Marsh, F. Sassi, V. L. Harvey, C. E. Randall, L. Emmons, J. F. Lamarque, P. Hess, J. J. Orlando, X. X. Tie, W. Randel, L. L. Pan, A. Gettelman, C. Granier, T. Diehl, U. Niemeier, and A. J. Simmons. Sensitivity of chemical tracers to meteorological parameters in the MOZART-3 chemical transport model. Journal of Geophysical Research, 112(D20302):1–24, October 2007. ISSN 0148-0227. doi: 10.1029/2006JD007879. URL http://doi.wiley.com/10.1029/2006JD007879.

P. K. Kitanidis and J. Lee. Principal Component Geostatistical Approach for large-dimensional inverse problems. Water Resources Research, 50(7):5428–5443, July 2014. ISSN 00431397. doi: 10.1002/2013WR014630. URL http://doi.wiley.com/10.1002/2013WR014630.

D. Koch and A. D. Del Genio. Black carbon semi-direct effects on cloud cover: review and synthesis. Atmospheric Chemistry and Physics, 10(16):7685–7696, August 2010. ISSN 1680-7324. doi: 10.5194/acp-10-7685-2010. URL http://www.atmos-chem-phys.net/10/7685/2010/.

Y. Kondo, L. Sahu, N. Moteki, F. Khan, N. Takegawa, X. Liu, M. Koike, and T. Miyakawa. Consistency and Traceability of Black Carbon Measurements Made by Laser-Induced Incandescence, Thermal-Optical Transmittance, and Filter-Based Photo-Absorption Techniques. Aerosol Science and Technology, 45(2):295–312, February 2011. ISSN 0278-6826, 1521-7388. doi: 10.1080/02786826.2010.533215. URL http://www.tandfonline.com/doi/abs/10.1080/02786826.2010.533215.

Daniel Krewski, Michael Jerrett, Richard T. Burnett, Renjun Ma, Edward Hughes, Yuanli Shi, Michelle C. Turner, C. Arden Pope III, George Thurston, and Eugenia E. Calle. Extended Follow-Up and Spatial Analysis of the American Cancer Society Study Linking Particulate Air Pollution and Mortality. HEI Research Report 140, Health Effects Institute, Boston, MA, May 2009. URL http://www.healtheffects.org/Pubs/RR140-Krewski.pdf.

Thomas Lauvaux, Natasha L. Miles, Aijun Deng, Scott J. Richardson, Maria O. Cambaliza, Kenneth J. Davis, Brian Gaudet, Kevin R. Gurney, Jianhua Huang, Darragh O'Keefe, Yang Song, Anna Karion, Tomohiro Oda, Risa Patarasuk, Igor Razlivanov, Daniel Sarmiento, Paul Shepson, Colm Sweeney, Jocelyn Turnbull, and Kai Wu. High-resolution atmospheric inversion of urban CO2 emissions during the dormant season of the Indianapolis Flux Experiment (INFLUX). Journal of Geophysical Research: Atmospheres, pages n/a–n/a, 2016. ISSN 2169-8996. doi: 10.1002/2015JD024473. URL http://dx.doi.org/10.1002/2015JD024473.

A. S. Lawless, S. Gratton, and N. K. Nichols. Approximate iterative methods for variational data assimilation. International Journal for Numerical Methods in Fluids, 47(10-11):1129–1135, 2005. ISSN 1097-0363. doi: 10.1002/fld.851. URL http://dx.doi.org/10.1002/fld.851.

P.F. Levelt, G.H.J. van den Oord, M.R. Dobber, A. Malkki, Huib Visser, J. de Vries, P. Stammes, J.O.V. Lundell, and H. Saari. The ozone monitoring instrument. Geoscience and Remote Sensing, IEEE Transactions on, 44(5):1093–1101, May 2006. ISSN 0196-2892. doi: 10.1109/TGRS.2006.872333.

Xin Liu, Zaiwen Wen, and Yin Zhang. An Efficient Gauss–Newton Algorithm for Symmetric Low-Rank Product Matrix Approximations. SIAM Journal on Optimization, 25(3):1571–1608, January 2015. ISSN 1052-6234, 1095-7189. doi: 10.1137/140971464. URL http://epubs.siam.org/doi/10.1137/140971464.

Zhiquan Liu, Quanhua Liu, Hui-Chuan Lin, Craig S. Schwartz, Yen-Huei Lee, and Tijian Wang. Three-dimensional variational assimilation of MODIS aerosol optical depth: Implementation and application to a dust storm over East Asia: AOD DATA ASSIMILATION. Journal of Geophysical Research: Atmospheres, 116(D23206):1–19, December 2011. ISSN 01480227. doi: 10.1029/2011JD016159. URL http://doi.wiley.com/10.1029/2011JD016159.

U. Lohmann and J. Feichter. Global indirect aerosol effects: a review. Atmospheric Chemistry and Physics, 5(3):715–737, 2005. doi: 10.5194/acp-5-715-2005. URL http://www.atmos-chem-phys.net/5/715/2005/.

Andrew C. Lorenc. The potential of the ensemble Kalman filter for NWP—a comparison with 4d-Var. Quarterly Journal of the Royal Meteorological Society, 129(595):3183–3203, 2003. ISSN 1477-870X. doi: 10.1256/qj.02.132. URL http://dx.doi.org/10.1256/qj.02.132.

P.A. Makar, W. Gong, J. Milbrandt, C. Hogrefe, Y. Zhang, G. Curci, R. Žabkar, U. Im, A. Balzarini, R. Baró, R. Bianconi, P. Cheung, R. Forkel, S. Gravel, M. Hirtl, L. Honzak, A. Hou, P. Jiménez-Guerrero, M. Langer, M.D. Moran, B. Pabla, J.L. Pérez, G. Pirovano, R. San José, P. Tuccella, J. Werhahn, J. Zhang, and S. Galmarini. Feedbacks between air pollution and weather, Part 1: Effects on weather. Atmospheric Environment, 115:442–469, August 2015. ISSN 1352-2310. doi: 10.1016/j.atmosenv.2014.12.003. URL http://www.sciencedirect.com/science/article/pii/S1352231014009510.

William C. Malm, James F. Sisler, Dale Huffman, Robert A. Eldred, and Thomas A. Cahill. Spatial and seasonal trends in particle concentration and optical extinction in the United States. Journal of Geophysical Research: Atmospheres, 99(D1):1347–1370, 1994. ISSN 2156-2202. doi: 10.1029/93JD02916. URL http://dx.doi.org/10.1029/93JD02916.

J. Mandel, E. Bergou, S. Gürol, S. Gratton, and I. Kasanický. Hybrid Levenberg–Marquardt and weak-constraint ensemble Kalman smoother method. Nonlinear Processes in Geophysics, 23(2):59–73, March 2016. ISSN 1607-7946. doi: 10.5194/npg-23-59-2016. URL http://www.nonlin-processes-geophys.net/23/59/2016/.

Y. H. Mao, Q. B. Li, L. Zhang, Y. Chen, J. T. Randerson, D. Chen, and K. N. Liou. Biomass burning contribution to black carbon in the Western United States Mountain Ranges. Atmospheric Chemistry and Physics, 11(21):11253–11266, November 2011. ISSN 1680-7324. doi: 10.5194/acp-11-11253-2011. URL http://www.atmos-chem-phys.net/11/11253/2011/.

Y. H. Mao, Q. B. Li, D. K. Henze, Z. Jiang, D. B. A. Jones, M. Kopacz, C. He, L. Qi, M. Gao, W.-M. Hao, and K.-N. Liou. Estimates of black carbon emissions in the western United States using the GEOS-Chem adjoint model. Atmospheric Chemistry and Physics, 15(13):7685–7702, 2015. doi: 10.5194/acp-15-7685-2015. URL http://www.atmos-chem-phys.net/15/7685/2015/.

G.I. Marchuk. Numerical solution of the problems of the dynamics of the atmosphere and ocean. Gitrometeoizadt, Leningrad, 1974.

Per-Gunnar Martinsson, Vladimir Rokhlin, and Mark Tygert. A randomized algorithm for the decomposition of matrices. Applied and Computational Harmonic Analysis, 30(1):47 – 68, 2011. ISSN 1063-5203. doi: http://dx.doi.org/10.1016/j.acha.2010.02.003. URL http://www.sciencedirect.com/science/article/pii/S1063520310000242.

Brian C. McDonald, Allen H. Goldstein, and Robert A. Harley. Long-Term Trends in California Mobile Source Emissions and Ambient Concentrations of Black Carbon and Organic Aerosol. Environmental Science & Technology, 49(8):5178–5188, April 2015. ISSN 0013-936X, 1520-5851. doi: 10.1021/es505912b. URL http://pubs.acs.org/doi/abs/10.1021/es505912b.

Jan Fokke Meirink, Peter Bergamaschi, and Maarten C. Krol. Four-dimensional variational data assimilation for inverse modelling of atmospheric methane emissions: method and comparison with synthesis inversion. Atmospheric chemistry and physics, 8(21):6341–6353, 2008. URL http://www.atmos-chem-phys.net/8/6341/.

J.-J. Morcrette, O. Boucher, L. Jones, D. Salmond, P. Bechtold, A. Beljaars, A. Benedetti, A. Bonet, J. W. Kaiser, M. Razinger, M. Schulz, S. Serrar, A. J. Simmons, M. Sofiev, M. Suttie, A. M. Tompkins, and A. Untch. Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: Forward modeling. Journal of Geophysical Research, 114(D06206):1–17, March 2009. ISSN 0148-0227. doi: 10.1029/2008JD011235. URL http://doi.wiley.com/10.1029/2008JD011235.

J.-F. Müller and T. Stavrakou. Inversion of CO and NO x emissions using the adjoint of the IMAGES model. Atmospheric Chemistry and Physics, 5(5):1157–1186, 2005. URL http://www.atmos-chem-phys.net/5/1157/2005/acp-5-1157-2005.html.

G. Myhre, D. Shindell, F.M. Breon, W. Collins, J. Fuglestvedt, J. Huang, D. Koch, J.F. Lamarque, D. Lee, B. Mendoza, T. Nakajima, A. Robock, G. Stephens, T. Takemura, and H. Zhang. Anthropogenic and Natural Radiative Forcing. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)], pages 129–234. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

NASA. MCD14ml MODIS Active Fire Detections. URL https://earthdata.nasa.gov/active-fire-data#tab-content-6.

M. Pagowski, G. A. Grell, S. A. McKeen, S. E. Peckham, and D. Devenyi. Three-dimensional variational data assimilation of ozone and fine particulate matter observations: some results using the Weather Research and Forecasting-Chemistry model and Grid-point Statistical Interpolation. Quarterly Journal of the Royal Meteorological Society, 136(653):2013–2024, October 2010. ISSN 00359009. doi: 10.1002/qj.700. URL http://doi.wiley.com/10.1002/qj.700.

Mariusz Pagowski and Georg A. Grell. Experiments with the assimilation of fine aerosols using an ensemble Kalman filter. Journal of Geophysical Research: Atmospheres, 117(D21302):1–15, 2012. ISSN 2156-2202. doi: 10.1029/2012JD018333. URL http://dx.doi.org/10.1029/2012JD018333.

J. Peischl, T. B. Ryerson, J. Brioude, K. C. Aikin, A. E. Andrews, E. Atlas, D. Blake, B. C. Daube, J. A. de Gouw, E. Dlugokencky, G. J. Frost, D. R. Gentner, J. B. Gilman, A. H. Goldstein, R. A. Harley, J. S. Holloway, J. Kofler, W. C. Kuster, P. M. Lang, P. C. Novelli, G. W. Santoni, M. Trainer, S. C. Wofsy, and D. D. Parrish. Quantifying sources of methane using light alkanes in the Los Angeles basin, California. Journal of Geophysical Research: Atmospheres, 118(10): 4974–4990, 2013. ISSN 2169-8996. doi: 10.1002/jgrd.50413. URL http://dx.doi.org/10.1002/jgrd.50413.

V. Penenko and N.N. Obraztsov. A variational initialization method for the fields of the meteorological elements. Meteorol. Gidrol., 11(1), 1976.

Jonathan E. Pleim. A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part II: Application and Evaluation in a Mesoscale Meteorological Model. Journal of Applied Meteorology and Climatology, 46(9):1396–1409, September 2007a. ISSN 1558-8424, 1558-8432. doi: 10.1175/JAM2534.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JAM2534.1.

Jonathan E. Pleim. A Combined Local and Nonlocal Closure Model for the Atmospheric Boundary Layer. Part I: Model Description and Testing. Journal of Applied Meteorology and Climatology, 46(9):1383–1395, September 2007b. ISSN 1558-8424, 1558-8432. doi: 10.1175/JAM2539.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JAM2539.1.

Jonathan E. Pleim. A Simple, Efficient Solution of Flux–Profile Relationships in the Atmospheric Surface Layer. Journal of Applied Meteorology and Climatology, 45(2):341–347, February 2006. ISSN 1558-8424. doi: 10.1175/JAM2339.1. URL http://dx.doi.org/10.1175/JAM2339.1.

Jonathan E. Pleim and Robert Gilliam. An Indirect Data Assimilation Scheme for Deep Soil Temperature in the Pleim–Xiu Land Surface Model. Journal of Applied Meteorology and Climatology, 48(7):1362–1376, July 2009. ISSN 1558-8424, 1558-8432. doi: 10.1175/2009JAMC2053.1. URL http://journals.ametsoc.org/doi/abs/10.1175/2009JAMC2053.1.

Jonathan E. Pleim and Aijun Xiu. Development of a land surface model. Part II: Data assimilation. Journal of Applied Meteorology, 42(12):1811–1822, 2003. URL http://journals.ametsoc.org/doi/full/10.1175/1520-0450(2003)042%3C1811:DOALSM%3E2.0.CO%3B2.

R. J. Purser and H.-L. Huang. Estimating Effective Data Density in a Satellite Retrieval or an Objective Analysis. Journal of Applied Meteorology, 32(6):1092–1107, 1993. doi: 10.1175/1520-0450(1993)032⟨1092:EEDDIA⟩2.0.CO;2.

Vishwas Rao and Adrian Sandu. A time-parallel approach to strong-constraint four-dimensional variational data assimilation. Journal of Computational Physics, 313:583 – 593, 2016. ISSN 0021-9991. doi: http://dx.doi.org/10.1016/j.jcp.2016.02.040. URL http://www.sciencedirect.com/science/article/pii/S0021999116001042.

Adam Reff, Prakash V. Bhave, Heather Simon, Thompson G. Pace, George A. Pouliot, J. David Mobley, and Marc Houyoux. Emissions Inventory of PM2.5 Trace Elements across the United States. Environmental Science & Technology, 43(15):5790–5796, August 2009. ISSN 0013-936X, 1520-5851. doi: 10.1021/es802930x. URL http://pubs.acs.org/doi/abs/10.1021/es802930x.

Jeffrey S. Reid, Edward J. Hyer, Elaine M. Prins, Douglas L. Westphal, Jianglong Zhang, Jun Wang, Sundar A. Christopher, Cynthia A. Curtis, Christopher C. Schmidt, Daniel P. Eleuterio,

Kim A. Richardson, and Jay P. Hoffman. Global Monitoring and Forecasting of Biomass-Burning Smoke: Description of and Lessons From the Fire Locating and Modeling of Burning Emissions (FLAMBE) Program. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2(3):144–162, September 2009. ISSN 1939-1404. doi: 10.1109/JSTARS.2009.2027443. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5208306.

Clive D. Rodgers. Information content and optimization of high-spectral-resolution measurements. In Proc. SPIE, volume 2830, pages 136–147, 1996. doi: 10.1117/12.256110.

Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A Randomized Algorithm for Principal Component Analysis. SIAM Journal on Matrix Analysis and Applications, 31(3):1100–1124, January 2010. ISSN 0895-4798, 1095-7162. doi: 10.1137/080736417. URL http://epubs.siam.org/doi/abs/10.1137/080736417.

Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled Newton methods I: globally convergent algorithms. arXiv preprint arXiv:1601.04737, 2016. URL http://arxiv.org/abs/1601.04737.

L. K. Sahu, Y. Kondo, N. Moteki, N. Takegawa, Y. Zhao, M. J. Cubison, J. L. Jimenez, S. Vay, G. S. Diskin, A. Wisthaler, T. Mikoviny, L. G. Huey, A. J. Weinheimer, and D. J. Knapp. Emission characteristics of black carbon in anthropogenic and biomass burning plumes over California during ARCTAS-CARB 2008. Journal of Geophysical Research, 117(D16302):1–20, August 2012. ISSN 0148-0227. doi: 10.1029/2011JD017401. URL http://doi.wiley.com/10.1029/2011JD017401.

P. E. Saide, S. N. Spak, G. R. Carmichael, M. A. Mena-Carrasco, Q. Yang, S. Howell, D. C. Leon, J. R. Snider, A. R. Bandy, J. L. Collett, K. B. Benedict, S. P. de Szoeke, L. N. Hawkins, G. Allen, I. Crawford, J. Crosier, and S. R. Springston. Evaluating WRF-Chem aerosol indirect effects in Southeast Pacific marine stratocumulus during VOCALS-REx. Atmospheric Chemistry and Physics, 12(6):3045–3064, 2012a. doi: 10.5194/acp-12-3045-2012. URL http://www.atmos-chem-phys.net/12/3045/2012/.

P. E. Saide, G. R. Carmichael, Z. Liu, C. S. Schwartz, H. C. Lin, A. M. da Silva, and E. Hyer. Aerosol optical depth assimilation for a size-resolved sectional model: impacts of observationally constrained, multi-wavelength and fine mode retrievals on regional scale analyses and forecasts. Atmospheric Chemistry and Physics, 13(20):10425–10444, October 2013. ISSN 1680-7324. doi: 10.5194/acp-13-10425-2013. URL http://www.atmos-chem-phys.net/13/10425/2013/.

P. E. Saide, S. N. Spak, R. B. Pierce, J. A. Otkin, T. K. Schaack, A. K. Heidinger, A. M. da Silva, M. Kacenelenbogen, J. Redemann, and G. R. Carmichael. Central American biomass burning smoke can increase tornado severity in the U.S. Geophysical Research Letters, 42(3):956–965, 2015a. ISSN 1944-8007. doi: 10.1002/2014GL062826. URL http://dx.doi.org/10.1002/2014GL062826.

Pablo E. Saide, Gregory R. Carmichael, Scott N. Spak, Patrick Minnis, and J. Kirk Ayers. Improving aerosol distributions below clouds by assimilating satellite-retrieved cloud droplet number. Proceedings of the National Academy of Sciences, 109(30):11939–11943, 2012b. URL http://www.pnas.org/content/109/30/11939.short.

Pablo E. Saide, David A. Peterson, Arlindo da Silva, Bruce Anderson, Luke D. Ziemba, Glenn Diskin, Glen Sachse, Johnathan Hair, Carolyn Butler, Marta Fenn, Jose L. Jimenez, Pedro Campuzano-Jost, Anne E. Perring, Joshua P. Schwarz, Milos Z. Markovic, Phil Russell, Jens Redemann, Yohei Shinozuka, David G. Streets, Fang Yan, Jack Dibb, Robert Yokelson, O. Brian Toon, Edward Hyer, and Gregory R. Carmichael. Revealing important nocturnal and day-to-day variations in fire smoke emissions through a multiplatform inversion. Geophysical Research Letters, 42(9):3609–3618, 2015b. ISSN 1944-8007. doi: 10.1002/2015GL063737. URL http://dx.doi.org/10.1002/2015GL063737.

B. H. Samset, G. Myhre, M. Schulz, Y. Balkanski, S. Bauer, T. K. Berntsen, H. Bian, N. Bellouin, T. Diehl, R. C. Easter, S. J. Ghan, T. Iversen, S. Kinne, A. Kirkevåg, J.-F. Lamarque, G. Lin, X. Liu, J. E. Penner, O. Seland, R. B. Skeie, P. Stier, T. Takemura, K. Tsigaridis, and K. Zhang. Black carbon vertical profiles strongly affect its radiative forcing uncertainty. Atmospheric Chemistry and Physics, 13(5):2423–2434, March 2013. ISSN 1680-7324. doi: 10.5194/acp-13-2423-2013. URL http://www.atmos-chem-phys.net/13/2423/2013/.

Adrian Sandu, Dacian N. Daescu, Gregory R. Carmichael, and Tianfeng Chai. Adjoint sensitivity analysis of regional air quality models. Journal of Computational Physics, 204(1):222–252, March 2005. ISSN 00219991. doi: 10.1016/j.jcp.2004.10.011. URL http://linkinghub.elsevier.com/retrieve/pii/S0021999104004140.

Yoshikazu Sasaki. Some basic formalisms in numerical variational analysis. Mon. Wea. Rev., 98(12): 875–883, December 1970. ISSN 0027-0644. doi: 10.1175/1520-0493(1970)098⟨0875:SBFINV⟩2.3.CO;2. URL http://dx.doi.org/10.1175/1520-0493(1970)098<0875:SBFINV>2.3.CO;2.

M. Schulz, C. Textor, S. Kinne, Y. Balkanski, S. Bauer, T. Berntsen, T. Berglen, O. Boucher, F. Dentener, S. Guibert, I. S. A. Isaksen, T. Iversen, D. Koch, A. Kirkevåg, X. Liu, V. Montanaro, G. Myhre, J. E. Penner, G. Pitari, S. Reddy, O. Seland, P. Stier, and T. Takemura. Radiative forcing by aerosols as derived from the AeroCom present-day and pre-industrial simulations. Atmospheric Chemistry and Physics, 6(12):5225–5246, 2006. doi: 10.5194/acp-6-5225-2006. URL http://www.atmos-chem-phys.net/6/5225/2006/.

Craig S. Schwartz, Zhiquan Liu, Hui-Chuan Lin, and Stuart A. McKeen. Simultaneous three-dimensional variational assimilation of surface fine particulate matter and MODIS aerosol optical depth. Journal of Geophysical Research, 117(D13202):1–22, July 2012. ISSN 0148-0227. doi: 10.1029/2011JD017383. URL http://doi.wiley.com/10.1029/2011JD017383.

Craig S. Schwartz, Zhiquan Liu, Hui-Chuan Lin, and Jeffrey D. Cetola. Assimilating aerosol observations with a "hybrid" variational-ensemble data assimilation system. Journal of Geophysical Research: Atmospheres, 119(7):4043–4069, 2014. ISSN 2169-8996. doi: 10.1002/2013JD020937. URL http://dx.doi.org/10.1002/2013JD020937.

Joel Schwartz, Brent Coull, Francine Laden, and Louise Ryan. The Effect of Dose and Timing of Dose on the Association between Airborne Particles and Survival. Environmental Health Perspectives, 116(1):64–69, October 2007. ISSN 0091-6765. doi: 10.1289/ehp.9955. URL http://www.ehponline.org/ambra-doi-resolver/10.1289/ehp.9955.

William C Skamarock, Joseph B Klemp, Jimy Dudhia, David O Gill, Dale M Barker, Michael G Duda, Xiang-Yu Huang, Wei Wang, and Jordan G Powers. A description of the advanced research

WRF version 3. Technical Report NCAR/TN-475+STR, DTIC Document, June 2008. URL http://nldr.library.ucar.edu/repository/collections/TECH-NOTE-000-000-000-855.

A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, and Y. Marzouk. Optimal Low-rank Approximations of Bayesian Linear Inverse Problems. SIAM J. Sci. Comput., 37(6):A2451–A2487, January 2015. ISSN 1064-8275. doi: 10.1137/140977308. URL http://dx.doi.org/10.1137/140977308.

D. V. Spracklen, L. J. Mickley, J. A. Logan, R. C. Hudman, R. Yevich, M. D. Flannigan, and A. L. Westerling. Impacts of climate change from 2000 to 2050 on wildfire activity and carbonaceous aerosol concentrations in the western United States. Journal of Geophysical Research, 114(D20), October 2009. ISSN 0148-0227. doi: 10.1029/2008JD010966. URL http://doi.wiley.com/10.1029/2008JD010966.

D. G. Streets, T. C. Bond, G. R. Carmichael, S. D. Fernandes, Q. Fu, D. He, Z. Klimont, S. M. Nelson, N. Y. Tsai, M. Q. Wang, J.-H. Woo, and K. F. Yarber. An inventory of gaseous and primary aerosol emissions in Asia in the year 2000. Journal of Geophysical Research: Atmospheres, 108(D21):GTE30/1–GTE30/23, 2003. ISSN 2156-2202. doi: 10.1029/2002JD003093. URL http://dx.doi.org/10.1029/2002JD003093.

Riku Suutari, Markus Amann, Janusz Cofala, Zbiegniew Klimont, Maximilian Posch, and Wolfgang Schöpp. From economic activities to ecosystem protection in Europe – An uncertainty analysis of two scenarios of the RAINS integrated assessment model. Technical Report CIAM/CCE Rep. 1/2001, Int. Inst. for Appl. Syst. Anal., Laxenburg, Austria, August 2001.

C. Textor, M. Schulz, S. Guibert, S. Kinne, Y. Balkanski, S. Bauer, T. Berntsen, T. Berglen, O. Boucher, M. Chin, F. Dentener, T. Diehl, R. Easter, H. Feichter, D. Fillmore, S. Ghan, P. Ginoux, S. Gong, A. Grini, J. Hendricks, L. Horowitz, P. Huang, I. Isaksen, I. Iversen, S. Kloster, D. Koch, A. Kirkevåg, J. E. Kristjansson, M. Krol, A. Lauer, J. F. Lamarque, X. Liu, V. Montanaro, G. Myhre, J. Penner, G. Pitari, S. Reddy, O. Seland, P. Stier, T. Takemura, and X. Tie. Analysis and quantification of the diversities of aerosol life cycles within AeroCom. Atmospheric Chemistry and Physics, 6(7):1777–1813, 2006. doi: 10.5194/acp-6-1777-2006. URL http://www.atmos-chem-phys.net/6/1777/2006/.

William Carlisle Thacker. The role of the Hessian matrix in fitting models to measurements. Journal of Geophysical Research: Oceans, 94(C5):6177–6196, 1989. ISSN 2156-2202. doi: 10.1029/JC094iC05p06177. URL http://dx.doi.org/10.1029/JC094iC05p06177.

Yannick Trémolet. Accounting for an imperfect model in 4d-Var. Quarterly Journal of the Royal Meteorological Society, 132(621):2483–2504, 2006. ISSN 1477-870X. doi: 10.1256/qj.05.224. URL http://dx.doi.org/10.1256/qj.05.224.

J. Tshimanga, S. Gratton, A. T. Weaver, and A. Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. Quarterly Journal of the Royal Meteorological Society, 134(632):751–769, April 2008. ISSN 00359009, 1477870X. doi: 10.1002/qj.228. URL http://doi.wiley.com/10.1002/qj.228.

A. J. Turner, D. K. Henze, R. V. Martin, and A. Hakami. The spatial extent of source influences on modeled column concentrations of short-lived species: Modeled Source Influences. Geophysical Research Letters, 39(L12806):1–5, June 2012. ISSN 00948276. doi: 10.1029/2012GL051832. URL http://doi.wiley.com/10.1029/2012GL051832.

Matthew D. Turner, Daven K. Henze, Amir Hakami, Shunliu Zhao, Jaroslav Resler, Gregory R. Carmichael, Charles O. Stanier, Jaemeen Baek, Adrian Sandu, Armistead G. Russell, Athanasios Nenes, Gill-Ran Jeong, Shannon L. Capps, Peter B. Percell, Rob W. Pinder, Sergey L. Napelenok, Jesse O. Bash, and Tianfeng Chai. Differences Between Magnitudes and Health Impacts of BC Emissions Across the United States Using 12 km Scale Seasonal Source Apportionment. Environmental Science & Technology, 49(7):4362–4371, 2015. doi: 10.1021/es505968b. URL http://dx.doi.org/10.1021/es505968b.

S. Twomey. The Influence of Pollution on the Shortwave Albedo of Clouds. Journal of the Atmospheric Sciences, 34(7):1149–1152, July 1977. ISSN 0022-4928. doi: 10.1175/1520-0469(1977)034⟨1149:TIOPOT⟩2.0.CO;2. URL http://dx.doi.org/10.1175/1520-0469(1977)034<1149:TIOPOT>2.0.CO;2.

UC-Davis. Interagency Monitoring of Protected Visual Environments Quality Assurance Project Plan. Technical report, March 2002. URL http://vista.cira.colostate.edu/improve/Publications/QA_QC/IMPROVE_QAPP_R0.pdf.

G. R. van der Werf, J. T. Randerson, L. Giglio, G. J. Collatz, M. Mu, P. S. Kasibhatla, D. C. Morton, R. S. DeFries, Y. Jin, and T. T. van Leeuwen. Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997-2009). Atmospheric Chemistry and Physics, 10(23):11707–11735, 2010. doi: 10.5194/acp-10-11707-2010. URL http://www.atmos-chem-phys.net/10/11707/2010/.

B. Vogel, H. Vogel, D. Bäumer, M. Bangert, K. Lundgren, R. Rinke, and T. Stanelle. The comprehensive model system COSMO-ART–Radiative impact of aerosol on the state of the atmosphere on the regional scale. Atmospheric Chemistry and Physics, 9(22):8661–8680, 2009. URL http://www.atmos-chem-phys.net/9/8661/2009/.

Grace Wahba. Design Criteria and Eigensequence Plots for Satellite-Computed Tomography. Journal of Atmospheric and Oceanic Technology, 2(2):125–132, 1985. doi: 10.1175/1520-0426(1985)002⟨0125:DCAEPF⟩2.0.CO;2. URL http://dx.doi.org/10.1175/1520-0426(1985)002<0125:DCAEPF>2.0.CO;2.

Jun Wang, Xiaoguang Xu, Daven K. Henze, Jing Zeng, Qiang Ji, Si-Chee Tsay, and Jianping Huang. Top-down estimate of dust emissions through integration of MODIS and MISR aerosol retrievals with the GEOS-Chem adjoint model. Geophysical Research Letters, 39(L08802):1–6, 2012. ISSN 1944-8007. doi: 10.1029/2012GL051136. URL http://dx.doi.org/10.1029/2012GL051136.

A. T. Weaver, C. Deltel, E. Machu, S. Ricci, and N. Daget. A multivariate balance operator for variational ocean data assimilation. Quarterly Journal of the Royal Meteorological Society, 131(613):3605–3625, October 2005. ISSN 00359009, 1477870X. doi: 10.1256/qj.05.119. URL http://doi.wiley.com/10.1256/qj.05.119.

K. J. Wecht, D. J. Jacob, M. P. Sulprizio, G. W. Santoni, S. C. Wofsy, R. Parker, H. Bösch, and J. Worden. Spatially resolving methane emissions in California: constraints from the CalNex aircraft campaign and from present (GOSAT, TES) and future (TROPOMI, geostationary) satellite observations. Atmospheric Chemistry and Physics, 14(15):8173–8184, 2014. doi: 10.5194/acp-14-8173-2014. URL http://www.atmos-chem-phys.net/14/8173/2014/.

M. L. Wesely. Parameterization of surface resistances to gaseous dry deposition in regional-scale numerical models. Atmospheric Environment (1967), 23(6):1293 – 1304, 1989. ISSN 0004-6981. doi: http://dx.doi.org/10.1016/0004-6981(89)90153-4. URL http://www.sciencedirect.com/science/article/pii/0004698189901534.

C. Wiedinmyer, S. K. Akagi, R. J. Yokelson, L. K. Emmons, J. A. Al-Saadi, J. J. Orlando, and A. J. Soja. The Fire INventory from NCAR (FINN): a high resolution global model to estimate the emissions from open burning. Geoscientific Model Development, 4(3):625–641, July 2011. ISSN 1991-9603. doi: 10.5194/gmd-4-625-2011. URL http://www.geosci-model-dev.net/4/625/2011/.

Christine Wiedinmyer, Brad Quayle, Chris Geron, Angie Belote, Don McKenzie, Xiaoyang Zhang, Susan O'Neill, and Kristina Klos Wynne. Estimating emissions from fires in North America for air quality modeling. Atmospheric Environment, 40(19):3419–3432, June 2006. ISSN 13522310. doi: 10.1016/j.atmosenv.2006.02.010. URL http://linkinghub.elsevier.com/retrieve/pii/S1352231006002032.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A fast randomized algorithm for the approximation of matrices. Applied and Computational Harmonic Analysis, 25(3):335 – 366, 2008. ISSN 1063-5203. doi: http://dx.doi.org/10.1016/j.acha.2007.12.002. URL http://www.sciencedirect.com/science/article/pii/S1063520307001364.

WRAP. 2002 Fire Emission Inventory for the WRAP Region - Phase II. Technical Report Project No. 178-6, Prepared by Air Sciences, Inc., July 2005. URL http://www.wrapair.org/forums/fejf/tasks/FEJFtask7PhaseII.html.

Jun Wu, Arthur M. Winer, and Ralph J. Delfino. Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. Atmospheric Environment, 40(18):3333 – 3348, 2006. ISSN 1352-2310. doi: http://dx.doi.org/10.1016/j.atmosenv.2006.01.056. URL http://www.sciencedirect.com/science/article/pii/S135223100600197X.

Qingnong Xiao, Ying-Hwa Kuo, Zaizhong Ma, Wei Huang, Xiang-Yu Huang, Xiaoyan Zhang, Dale M. Barker, John Michalakes, and Jimy Dudhia. Application of an Adiabatic WRF Adjoint to the Investigation of the May 2004 McMurdo, Antarctica, Severe Wind Event. Monthly Weather Review, 136(10):3696–3713, October 2008. ISSN 0027-0644. doi: 10.1175/2008MWR2235.1. URL http://dx.doi.org/10.1175/2008MWR2235.1.

Aijun Xiu and Jonathan E. Pleim. Development of a land surface model. Part I: Application in a mesoscale meteorological model. Journal of Applied Meteorology, 40(2):192–209, 2001. URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0450(2001)040%3C0192:DOALSM%3E2.0.CO%3B2.

Eun-Gyeong Yang, Hyun Mee Kim, JinWoong Kim, and Jun Kyung Kay. Effect of Observation Network Design on Meteorological Forecasts of Asian Dust Events. Monthly Weather Review, 142(12):4679–4695, December 2014. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-14-00080.1. URL http://journals.ametsoc.org/doi/abs/10.1175/MWR-D-14-00080.1.

Q. Yang, W. I. Gustafson Jr., J. D. Fast, H. Wang, R. C. Easter, H. Morrison, Y.-N. Lee, E. G. Chapman, S. N. Spak, and M. A. Mena-Carrasco. Assessing regional scale predictions of aerosols, marine stratocumulus, and their interactions during VOCALS-REx using

WRF-Chem. Atmospheric Chemistry and Physics, 11(23):11951–11975, 2011. doi: 10.5194/acp-11-11951-2011. URL http://www.atmos-chem-phys.net/11/11951/2011/.

Tiffany L.B. Yelverton, Michael D. Hays, Brian K. Gullett, and William P. Linak. Black Carbon Measurements of Flame-Generated Soot as Determined by Optical, Thermal-Optical, Direct Absorption, and Laser Incandescence Methods. Environmental Engineering Science, 31(4):209–215, April 2014. ISSN 1092-8758, 1557-9018. doi: 10.1089/ees.2014.0038. URL http://online.liebertpub.com/doi/abs/10.1089/ees.2014.0038.

Feng Zhang, Jun Wang, Charles Ichoku, Edward J Hyer, Zhifeng Yang, Cui Ge, Shenjian Su, Xiaoyang Zhang, Shobha Kondragunta, Johannes W Kaiser, Christine Wiedinmyer, and Arlindo da Silva. Sensitivity of mesoscale modeling of smoke direct radiative effect to the emission inventory: a case study in northern sub-Saharan African region. Environmental Research Letters, 9(7): 075002, July 2014a. ISSN 1748-9326. doi: 10.1088/1748-9326/9/7/075002. URL http://stacks.iop.org/1748-9326/9/i=7/a=075002?key=crossref.c3a057c1cb5936f978ecde79c6fa7281.

Xiaoyang Zhang, Shobha Kondragunta, Jessica Ram, Christopher Schmidt, and Ho-Chun Huang. Near-real-time global biomass burning emissions product from geostationary satellite constellation: GLOBAL BIOMASS BURNING EMISSIONS. Journal of Geophysical Research: Atmospheres, 117(D14):n/a–n/a, July 2012. ISSN 01480227. doi: 10.1029/2012JD017459. URL http://doi.wiley.com/10.1029/2012JD017459.

Xin Zhang, Xiang-Yu Huang, and Ning Pan. Development of the Upgraded Tangent Linear and Adjoint of the Weather Research and Forecasting (WRF) Model. Journal of Atmospheric and Oceanic Technology, 30(6):1180–1188, February 2013. ISSN 0739-0572. doi: 10.1175/JTECH-D-12-00213.1. URL http://dx.doi.org/10.1175/JTECH-D-12-00213.1.

Xin Zhang, Xiang-Yu Huang, Jianyu Liu, Jonathan Poterjoy, Yonghui Weng, Fuqing Zhang, and Hongli Wang. Development of an Efficient Regional Four-Dimensional Variational Data Assimilation System for WRF. Journal of Atmospheric and Oceanic Technology, 31(12):2777–2794, December 2014b. ISSN 0739-0572, 1520-0426. doi: 10.1175/JTECH-D-13-00076.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JTECH-D-13-00076.1.

Appendix A

Relating DA and optimization formulations

The linear optimization in the inner loop solves a system

$$\begin{aligned} \min_{\hat{\boldsymbol{x}}} \quad & F(\hat{\boldsymbol{x}}) = \frac{1}{2}\hat{\boldsymbol{x}}^\top \mathbf{A}'\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^\top \boldsymbol{b} + c \\ & \mathbf{A}'\hat{\boldsymbol{x}} = \boldsymbol{b}. \end{aligned} \tag{A.1}$$

In our case, $\hat{\boldsymbol{x}} \equiv \delta\boldsymbol{v}^k$. The equivalence of Eq. 3.10 and Eq. A.1 is apparent in Tshimanga et al. (2008), who provide a notational translation between publications on DA and those on minimization algorithms and preconditioners. We repeat their translation to account for the differences in formulation of Eq. 3.10 and Eq. 5 in Tshimanga et al. (2008).

The process starts by considering Lawless et al. (2005) and Gratton et al. (2007), who show that incremental 4D-Var is equivalent to a truncated Gauss-Newton (TGN) optimization algorithm. The incremental 4D-Var cost function is condensed to:

$$\min_{\delta\boldsymbol{v}} \quad J(\delta\boldsymbol{v}) = \frac{1}{2}\mathbf{f}(\delta\boldsymbol{v})^\top \mathbf{f}(\delta\boldsymbol{v}), \tag{A.2}$$

where

$$\mathbf{f}\left(\delta\boldsymbol{v}^k\right) \equiv \begin{pmatrix} \delta\boldsymbol{v}^k - \boldsymbol{d}^{b,k-1} \\ \mathbf{R}^{-\frac{1}{2}}\left(\mathbf{G}^{k-1}\mathbf{U}\delta\boldsymbol{v}^k - \boldsymbol{d}^{o,k-1}\right) \end{pmatrix}. \tag{A.3}$$

This definition of $\mathbf{f}$ is what enables incremental 4D-Var to be characterized as TGN. The remainder of the derivation amounts to substitutions. GN approximates Newton's method in each quadratic minimization problem, k, to solve for the increment $\delta\boldsymbol{v}^k$ in the linearized system

$$\mathcal{H}_{\delta\boldsymbol{v}}\delta\boldsymbol{v}^k = -\nabla J. \tag{A.4}$$

This form is equivalent to multiplying Eq. 3.21 by the Hessian on both sides. In our case, the right-hand side is $\boldsymbol{b} \equiv -\nabla J = -{\mathbf{F}^{k-1}}^{\top}\mathbf{f}\left(\delta\boldsymbol{v}^{k-1}\right)$, where

$$\mathbf{f}\left(\delta\boldsymbol{v}^{k-1}\right) \equiv -\begin{pmatrix} \boldsymbol{d}^{b,k-1} \\ \mathbf{R}^{-\frac{1}{2}}\boldsymbol{d}^{o,k-1} \end{pmatrix}, \tag{A.5}$$

and

$$\mathbf{F}^{k-1} \equiv \nabla_{\delta\boldsymbol{v}^{k-1}}\mathbf{f}\big|_{\delta\boldsymbol{v}^{k-1}} = \begin{pmatrix} \mathbf{I}_n \\ \mathbf{R}^{-\frac{1}{2}}\mathbf{G}^{k-1}\mathbf{U} \end{pmatrix}. \tag{A.6}$$

$\mathbf{f}\left(\delta\boldsymbol{v}^{k-1}\right)$ and its Jacobian are fixed for each outer loop by the $k-1$ trajectory. Completing the GN algorithm, the Hessian ($\mathcal{H}_{\delta\boldsymbol{v}}$) is approximated by $\mathbf{A}' \equiv {\mathbf{F}^{k-1}}^{\top}\mathbf{F}^{k-1}$, after ignoring mixed partial derivatives of $\mathbf{f}$. The Hessian of Eq. A.2 matches that of Eq. 3.10, namely

$$\mathcal{H}_{\delta\boldsymbol{v}} = \mathbf{I}_n + \mathbf{U}^{\top}{\mathbf{G}^{k-1}}^{\top}\mathbf{R}^{-1}\mathbf{G}^{k-1}\mathbf{U}. \tag{A.7}$$

After substitutions, Eq. A.4 becomes

$${\mathbf{F}^{k-1}}^{\top}\mathbf{F}^{k-1}\delta\boldsymbol{v}^{k} = -{\mathbf{F}^{k-1}}^{\top}\mathbf{f}\left(\delta\boldsymbol{v}^{k-1}\right), \tag{A.8}$$

which expands to

$$\left(\mathbf{I}_n + \mathbf{U}^{\top}{\mathbf{G}^{k-1}}^{\top}\mathbf{R}^{-1}\mathbf{G}^{k-1}\mathbf{U}\right)\delta\boldsymbol{v}^{k} = \boldsymbol{d}^{b,k-1} + \mathbf{U}^{\top}{\mathbf{G}^{k-1}}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}. \tag{A.9}$$

Solving for $\delta\boldsymbol{v}^{k}$ gives the same update formula that would result from setting Eq. 3.20 equal to zero,

$$\delta\boldsymbol{v}^{k} = \left(\mathbf{I}_n + \mathbf{U}^{\top}{\mathbf{G}^{k-1}}^{\top}\mathbf{R}^{-1}\mathbf{G}^{k-1}\mathbf{U}\right)^{-1}\left(\boldsymbol{d}^{b,k-1} + \mathbf{U}^{\top}{\mathbf{G}^{k-1}}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right). \tag{A.10}$$

Thus, by defining $\mathbf{f}$ appropriately, the eqivalence between GN and incremental 4D-Var is verified.

Appendix B

Derivation of the truncated inverse Hessian

After l inner iterations, the Lanczos vectors form an orthogonal matrix, $\mathbf{Q}_l = [\hat{\boldsymbol{q}}_1, .., \hat{\boldsymbol{q}}_l]$, which satisfies

$$\mathcal{H}_{\delta \boldsymbol{v}} \mathbf{Q}_l = \mathbf{Q}_l \mathbf{T}_l. \tag{B.1}$$

The extremal eigenvalues of $\mathbf{T}_l$ are good approximations to $\mathcal{H}_{\delta \boldsymbol{v}}$'s extremal eigenvalues (Golub and Van Loan, 1996). $\mathbf{T}_l$ can be decomposed as

$$\mathbf{T}_l = \mathbf{W}_l \mathbf{\Lambda}_l \mathbf{W}_l^{-1}. \tag{B.2}$$

If we were to carry out the minimization for n steps, we would find all the Lanczos vectors, and would be able construct the full $\mathbf{T}$ and $\mathbf{Q}$ matrices. In that case, the orthogonal Lanczos vectors admit $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$. When combined with Eq. B.1,

$$\mathcal{H}_{\delta \boldsymbol{v}} = \mathbf{Q}\mathbf{W}\mathbf{\Lambda}\mathbf{W}^{-1}\mathbf{Q}^\top \tag{B.3}$$

Because the eigenvectors are orthonormal,

$$\mathcal{H}_{\delta \boldsymbol{v}} = (\mathbf{Q}\mathbf{W})\, \mathbf{\Lambda}\, (\mathbf{Q}\mathbf{W})^\top. \tag{B.4}$$

Thus, the eigenvectors of $\mathcal{H}_{\delta \boldsymbol{v}}$ are approximately equal to the normalized eigenvectors of $\mathbf{T}$, pre-multiplied by the matrix of Lanczos vectors, i.e.,

$$\mathcal{H}_{\delta \boldsymbol{v}} = \hat{\boldsymbol{\nu}} \mathbf{\Lambda} \hat{\boldsymbol{\nu}}^\top, \tag{B.5}$$

where the ${k_i}^{th}$ eigenvector of $\mathcal{H}_{\delta \boldsymbol{v}}$ is

$$\hat{\boldsymbol{\nu}}_{k_i} = \mathbf{Q}\hat{\boldsymbol{w}}_{k_i}. \tag{B.6}$$

The Hessian is constructed by

$$\mathcal{H}_{\delta \boldsymbol{v}} = \hat{\boldsymbol{\nu}}\boldsymbol{\Lambda}\hat{\boldsymbol{\nu}}^\top = \sum_{k_i=1}^{n} \lambda_{k_i}\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^\top. \tag{B.7}$$

Since the Hessian and its inverse have identical eigenvectors and reciprocal eigenvalues, the inverse is

$$[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1} = \sum_{k_i=1}^{n} \lambda_{k_i}^{-1}\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^\top. \tag{B.8}$$

Although this expression is usable, computational resource limitations require $l << n$. Truncating the sum yields a low rank estimate for the inverse, and for the posterior error which it estimates.

A more robust estimate of the posterior error is a low-rank update to the full-rank prior covariance, $\mathbf{B}$. To pursue that goal, first we return to the linear algebra formula, then add and subtract the identity matrix to get

$$\begin{aligned}
\mathcal{H}_{\delta \boldsymbol{v}} &= \mathbf{I} + \hat{\boldsymbol{\nu}}\boldsymbol{\Lambda}\hat{\boldsymbol{\nu}}^\top - \mathbf{I} \\
&= \mathbf{I} + \hat{\boldsymbol{\nu}}\boldsymbol{\Lambda}\hat{\boldsymbol{\nu}}^\top - \hat{\boldsymbol{\nu}}\mathbf{I}\hat{\boldsymbol{\nu}}^\top \\
&= \mathbf{I} + \sum_{k_i=1}^{n} \left(\lambda_{k_i} - 1\right)\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^\top
\end{aligned} \tag{B.9}$$

Now we repeat the truncation,

$$\mathcal{H}_{\delta \boldsymbol{v}} \approx \mathbf{I} + \sum_{k_i=1}^{l} \left(\lambda_{k_i} - 1\right)\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^\top, \tag{B.10}$$

where $\hat{\boldsymbol{\nu}}_{k_i}$ is constructed from the partial set of Lanczos vectors as

$$\hat{\boldsymbol{\nu}}_{k_i} = \mathbf{Q}_l\hat{\boldsymbol{w}}_{lk_i}. \tag{B.11}$$

Next we apply the Sherman-Morrison formula to recursively build the inverse for each term in the sum. Throughout, we will take advantage of the following two relationships for orthogonal vectors

$$\hat{\boldsymbol{\nu}}_j^\top\hat{\boldsymbol{\nu}}_j = 1$$

and

$$\hat{\boldsymbol{\nu}}_j^\top\hat{\boldsymbol{\nu}}_i = 0; \quad j \neq i.$$

Starting with the first term,

$$\begin{aligned}\mathbf{N}_1^{-1} &= \left[\mathbf{I} + (\lambda_1 - 1)\,\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top\right]^{-1} \\ &= \mathbf{I}^{-1} - \frac{\mathbf{I}^{-1}(\lambda_1 - 1)\,\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top\mathbf{I}^{-1}}{1 + (\lambda_1 - 1)\,\hat{\boldsymbol{\nu}}_1^\top\mathbf{I}^{-1}\hat{\boldsymbol{\nu}}_1} \\ &= \mathbf{I} - \frac{(\lambda_1 - 1)\,\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top}{\lambda_1} \\ &= \mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top.\end{aligned}$$

This result fits our desired proof. Now for the second term,

$$\begin{aligned}\mathbf{N}_2^{-1} &= \left\{\mathbf{N}_1 + (\lambda_2 - 1)\,\hat{\boldsymbol{\nu}}_2\hat{\boldsymbol{\nu}}_2^\top\right\}^{-1} \\ &= \mathbf{N}_1^{-1} - \frac{\mathbf{N}_1^{-1}(\lambda_2 - 1)\,\hat{\boldsymbol{\nu}}_2\hat{\boldsymbol{\nu}}_2^\top\mathbf{N}_1^{-1}}{1 + (\lambda_2 - 1)\,\hat{\boldsymbol{\nu}}_2^\top\mathbf{N}_1^{-1}\hat{\boldsymbol{\nu}}_2} \\ &= \mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top - \\ &\quad \frac{\left[\mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top\right](\lambda_2 - 1)\,\hat{\boldsymbol{\nu}}_2\hat{\boldsymbol{\nu}}_2^\top\left[\mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top\right]}{1 + (\lambda_2 - 1)\,\hat{\boldsymbol{\nu}}_2^\top\left[\mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top\right]\hat{\boldsymbol{\nu}}_2} \\ &= \mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top + \left(\lambda_2^{-1} - 1\right)\hat{\boldsymbol{\nu}}_2\hat{\boldsymbol{\nu}}_2^\top.\end{aligned}$$

The dot products of orthogonal vectors cancels all terms in the numerator and denominator except the ones multiplied by the identity matrix. The same simplification applies to each additional sum, where the full sum can be expressed as

$$\begin{aligned}\mathbf{N}_l^{-1} = &\mathbf{I} + \left(\lambda_1^{-1} - 1\right)\hat{\boldsymbol{\nu}}_1\hat{\boldsymbol{\nu}}_1^\top - \\ &\sum_{k_i=2}^{l} \frac{\left[\mathbf{I} + \sum_{r=1}^{k_i-1}\left(\lambda_r^{-1} - 1\right)\hat{\boldsymbol{\nu}}_r\hat{\boldsymbol{\nu}}_r^\top\right](\lambda_{k_i} - 1)\,\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^\top\left[\mathbf{I} + \sum_{r=1}^{k_i-1}\left(\lambda_r^{-1} - 1\right)\hat{\boldsymbol{\nu}}_r\hat{\boldsymbol{\nu}}_r^\top\right]}{1 + (\lambda_{k_i} - 1)\,\hat{\boldsymbol{\nu}}_{k_i}^\top\left[\mathbf{I} + \sum_{r=1}^{k_i-1}\left(\lambda_r^{-1} - 1\right)\hat{\boldsymbol{\nu}}_r\hat{\boldsymbol{\nu}}_r^\top\right]\hat{\boldsymbol{\nu}}_{k_i}}\end{aligned}$$

Here, again, all of the terms where $r \neq k_i$ cancel. What remains is similar to Eq. B.8, but slightly modified.

$$[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1} \approx \mathbf{I} + \sum_{k_i=1}^{l}\left(\lambda_{k_i}^{-1} - 1\right)\hat{\boldsymbol{\nu}}_{k_i}\hat{\boldsymbol{\nu}}_{k_i}^\top. \tag{B.12}$$

After a left-side multiplication by $\mathbf{U}$ and a right-side multiplication by $\mathbf{U}^\top$, we achieve the desired low rank update to $\mathbf{B}$ found in Eq. 3.26.

Appendix C

Inverse Hessian conversion

It is easy enough to derive the expression in Eq. 3.25,

$$\mathcal{H}_{\delta \boldsymbol{v}} = \mathbf{U}^\top \mathcal{H}_{\delta \boldsymbol{x}} \mathbf{U},$$

by starting from Eq. 3.24, then pre-multiplying by $\mathbf{U}^\top$ and post-multiplying by $\mathbf{U}$, and using the relationship $\mathbf{U}^\top \mathbf{B}^{-1} \mathbf{U} = \mathbf{I}$. If $\mathbf{U}$ and $\mathbf{U}^\top$ are invertible, then it is also easy to derive the expression

$$\mathcal{H}_{\delta \boldsymbol{x}}^{-1} = \mathbf{U} \mathcal{H}_{\delta \boldsymbol{v}}^{-1} \mathbf{U}^\top. \tag{C.1}$$

In the case that $\mathbf{U}$ is not invertible (e.g., singular), some relationship like Eq. C.1 must hold to be able to calculate the posterior covariance in $\boldsymbol{x}$ space, $\mathbf{P}^a$, as opposed to in $\boldsymbol{v}$ space, $\mathbf{P}_{\boldsymbol{v}}^a$.

To do so we start from the increment formulae in $\boldsymbol{x}$ space

$$\delta \boldsymbol{x}^k = -\mathcal{H}_{\delta \boldsymbol{x}}^{-1} \nabla_{\delta \boldsymbol{x}} J|_{\delta \boldsymbol{x}^k = \mathbf{0}}, \tag{C.2}$$

and $\boldsymbol{v}$ space

$$\delta \boldsymbol{v}^k = -\mathcal{H}_{\delta \boldsymbol{v}}^{-1} \nabla_{\delta \boldsymbol{v}} J|_{\delta \boldsymbol{v}^k = \mathbf{0}}. \tag{C.3}$$

Knowing that $\delta \boldsymbol{x}^k = \mathbf{U} \delta \boldsymbol{v}^k$, then the following equality holds

$$\mathcal{H}_{\delta \boldsymbol{x}}^{-1} \nabla_{\delta \boldsymbol{x}} J|_{\delta \boldsymbol{x}^k = \mathbf{0}} = \mathbf{U} \mathcal{H}_{\delta \boldsymbol{v}}^{-1} \nabla_{\delta \boldsymbol{v}} J|_{\delta \boldsymbol{v}^k = \mathbf{0}}. \tag{C.4}$$

Therefore,

$$\begin{aligned}\mathcal{H}_{\delta \boldsymbol{x}}^{-1}\left(\mathbf{B}^{-1}\sum_{k_o=1}^{k}\delta \boldsymbol{x}^{k^o}+\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right)&=\mathbf{U}\mathcal{H}_{\delta \boldsymbol{v}}^{-1}\left(\boldsymbol{d}^{b,k-1}+\mathbf{U}^{\top}\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right)\\ \mathcal{H}_{\delta \boldsymbol{x}}^{-1}\left(\mathbf{B}^{-1}\mathbf{U}\boldsymbol{d}^{b,k-1}+\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right)&=\mathbf{U}\mathcal{H}_{\delta \boldsymbol{v}}^{-1}\left(\boldsymbol{d}^{b,k-1}+\mathbf{U}^{\top}\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right).\end{aligned} \tag{C.5}$$

To confirm that Eq. C.1 is true, we substitute it into this relationship and simplify as follows

$$\begin{aligned}\mathbf{U}\mathcal{H}_{\delta \boldsymbol{v}}^{-1}\mathbf{U}^{\top}\left(\mathbf{B}^{-1}\mathbf{U}\boldsymbol{d}^{b,k-1}+\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right)&=\mathbf{U}\mathcal{H}_{\delta \boldsymbol{v}}^{-1}\left(\boldsymbol{d}^{b,k-1}+\mathbf{U}^{\top}\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right),\\ \mathbf{U}\mathcal{H}_{\delta \boldsymbol{v}}^{-1}\left(\boldsymbol{d}^{b,k-1}+\mathbf{U}^{\top}\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right)&=\mathbf{U}\mathcal{H}_{\delta \boldsymbol{v}}^{-1}\left(\boldsymbol{d}^{b,k-1}+\mathbf{U}^{\top}\mathbf{G}^{\top}\mathbf{R}^{-1}\boldsymbol{d}^{o,k-1}\right).\end{aligned} \tag{C.6}$$

The two sides are equivalent and we can use Eq. C.1, as well as

$$\mathbf{P}^{a}=\mathbf{U}\mathbf{P}_{\boldsymbol{v}}^{a}\mathbf{U}^{\top}, \tag{C.7}$$

where $\mathbf{P}^{a}=\mathcal{H}_{\delta \boldsymbol{x}}^{-1}$ and $\mathbf{P}_{\boldsymbol{v}}^{a}=\mathcal{H}_{\delta \boldsymbol{x}}^{-1}$.

Appendix D

RIOT Algorithms

Algorithm 3 RIOT-56

Require: $\hat{\mathbf{A}} \equiv \mathcal{H}_{\delta \boldsymbol{v},o} = \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-1} \mathbf{G} \mathbf{U} \in \mathbb{R}^{n \times n}$
and $\boldsymbol{\Omega} \in \mathbb{R}^{n \times N_{\text{app}}} \sim \mathcal{N}(0,1)$

Start with $x^{k=0} = x^b$, $\boldsymbol{v}^0 = \mathbf{0}$
for $k = 1, 2, \ldots, k_f$ **do**
  Calculate $J(x^{k-1})$ and
  store trajectory for $\mathbf{G}$ and $\mathbf{G}^\top$
  for all $k_i \in \{1, 2, \ldots, N_{\text{app}} + 1\}$ **do in parallel**
    if $k_i = N_{\text{app}}$ **then**
      $\boldsymbol{b} = -\nabla J|_{\delta \boldsymbol{v}=0} = -\boldsymbol{d}^{b,k-1} - \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-1} \boldsymbol{d}^{o,k-1}$
    else
      $\boldsymbol{\omega}_{k_i} = \boldsymbol{\Omega}(:, k_i)$
      $\boldsymbol{y}_{k_i} = \hat{\mathbf{A}} \boldsymbol{\omega}_{k_i}$
    end if
  end for
  $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{k_i}]$
  Calculate $\mathbf{Q} \in \mathbb{R}^{n \times N_{\text{app}}}$ from $\text{QR}(\mathbf{Y})$
  Solve for $\mathbf{K}$ in $\mathbf{K}\mathbf{Q}^\top \boldsymbol{\Omega} \approx \mathbf{Q}^\top \mathbf{Y}$.
  Form SVD, $\mathbf{K} = \mathbf{W} \boldsymbol{\Lambda}_1 \mathbf{Z}^\top$
  Form eigenmodes of $\mathcal{H}_{\delta \boldsymbol{v}}$: $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_1 + \mathbf{I}$ and $\mathbf{V} = \mathbf{Q}\mathbf{W}$
  $\delta \boldsymbol{v} = -\mathcal{H}^{-1}_{\delta \boldsymbol{v}} \nabla J|_{\delta \boldsymbol{v}=0}$
  $\boldsymbol{v}^k = \boldsymbol{v}^{k-1} + \delta \boldsymbol{v}$
  $\boldsymbol{x}^k = \boldsymbol{x}^{k-1} + \mathbf{U} \delta \boldsymbol{v}$
end for
$\mathbf{P}^a = \mathbf{U} \mathbf{P}^a_{\boldsymbol{v}} \mathbf{U}^\top$

Algorithm 4 RIOT-51

Require: $\hat{\mathbf{A}} \equiv \mathcal{H}^{\frac{1}{2}}_{\delta \boldsymbol{v},o} = \mathbf{R}^{-\frac{1}{2}} \mathbf{G} \mathbf{U} \in \mathbb{R}^{m \times n}$
and $\boldsymbol{\Omega} \in \mathbb{R}^{n \times N_{\text{app}}} \sim \mathcal{N}(0,1)$

Start with $x^{k=0} = x^b$, $\boldsymbol{v}^0 = \mathbf{0}$
for $k = 1, 2, \ldots, k_f$ **do**
  Calculate $J(x^{k-1})$ and
  store trajectory for $\mathbf{G}$ and $\mathbf{G}^\top$
  for all $k_i \in \{1, 2, \ldots, N_{\text{app}} + 1\}$ **do in parallel**
    if $k_i = (N_{\text{app}} + 1)$ **then**
      $\boldsymbol{b} = -\nabla J|_{\delta \boldsymbol{v}=0} = -\boldsymbol{d}^{b,k-1} - \mathbf{U}^\top \mathbf{G}^\top \mathbf{R}^{-1} \boldsymbol{d}^{o,k-1}$
    else
      $\boldsymbol{\omega}_{k_i} = \boldsymbol{\Omega}(:, k_i)$
      $\boldsymbol{y}_{k_i} = \hat{\mathbf{A}} \boldsymbol{\omega}_{k_i}$
    end if
  end for
  Calculate $\mathbf{Q} \in \mathbb{R}^{m \times N_{\text{app}}}$ from QR($\mathbf{Y}$)
  for all $k_i \in \{1, 2, \ldots, N_{\text{app}}\}$ **do in parallel**
    $\boldsymbol{q}_{k_i} = \mathbf{Q}(:, k_i)$
    $\mathbf{K}_1(:, k_i) = (\hat{\mathbf{A}}^\top \boldsymbol{q}_{k_i})^\top$.
  end for
  Form SVD, $\mathbf{K}_1 = \mathbf{W}\mathbf{S}\mathbf{Z}^\top$
  Form eigenmodes of $\mathcal{H}_{\delta \boldsymbol{v}}$: $\boldsymbol{\Lambda} = \mathbf{S}^2 + \mathbf{I}$ and $\mathbf{V} = \mathbf{Z}$
  $\delta \boldsymbol{v} = -\mathcal{H}_{\delta \boldsymbol{v}}^{-1} \nabla J|_{\delta \boldsymbol{v}=0}$
  $\boldsymbol{v}^k = \boldsymbol{v}^{k-1} + \delta \boldsymbol{v}$
  $\boldsymbol{x}^k = \boldsymbol{x}^{k-1} + \mathbf{U} \delta \boldsymbol{v}$
end for
$\mathbf{P}^a = \mathbf{U} \mathbf{P}_{\boldsymbol{v}}^a \mathbf{U}^\top$

Appendix E

LRA versus LRU increment and posterior covariance

We discussed the using $\hat{\boldsymbol{b}}$ as one of the basis vectors of the Hessian in Sec. 4.2.2.3, and in Sec. 4.3.1 we demonstrated that doing so removes any perceptible difference between the LRA and LRU analysis increments. In this appendix, we will show why that happens, and also discuss other effects of using $\hat{\boldsymbol{b}}$. First, we start with the inverse of the preconditioned Hessian in matrix-form from the LRA

$$[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1}_{\text{LRA}} \approx \mathbf{Q}\mathbf{W}\boldsymbol{\Lambda}^{-1}\mathbf{W}^{\top}\mathbf{Q}^{\top}, \tag{E.1}$$

and the LRU

$$[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1}_{\text{LRU}} \approx \mathbf{I}_n - \mathbf{Q}\mathbf{W}\left(\mathbf{I}_l - \boldsymbol{\Lambda}^{-1}\right)\mathbf{W}^{\top}\mathbf{Q}^{\top}. \tag{E.2}$$

For these approximations, $\mathbf{Q} \in \mathbb{R}^{n\times l}$ and $\mathbf{W} \in \mathbb{R}^{l\times l}$. The difference between the two approximations of the inverse Hessian is

$$\begin{aligned}
\Delta_{[\mathcal{H}_{\delta \boldsymbol{v}}]^{-1}} &= [\mathcal{H}_{\delta \boldsymbol{v}}]^{-1}_{\text{LRA}} - [\mathcal{H}_{\delta \boldsymbol{v}}]^{-1}_{\text{LRU}} \\
&= \mathbf{Q}\mathbf{W}\boldsymbol{\Lambda}^{-1}\mathbf{W}^{\top}\mathbf{Q}^{\top} - \left[\mathbf{I}_n - \mathbf{Q}\mathbf{W}\left(\mathbf{I}_l - \boldsymbol{\Lambda}^{-1}\right)\mathbf{W}^{\top}\mathbf{Q}^{\top}\right] \\
&= \mathbf{Q}\mathbf{W}\left(\boldsymbol{\Lambda}^{-1} + \mathbf{I}_l - \boldsymbol{\Lambda}^{-1}\right)\mathbf{W}^{\top}\mathbf{Q}^{\top} - \mathbf{I}_n \\
&= \mathbf{Q}\mathbf{W}\mathbf{W}^{\top}\mathbf{Q}^{\top} - \mathbf{I}_n \\
&= \mathbf{Q}\mathbf{I}_l\mathbf{Q}^{\top} - \mathbf{I}_n \\
&= \mathbf{Q}\mathbf{Q}^{\top} - \mathbf{I}_n.
\end{aligned} \tag{E.3}$$

Therefore, the difference between the LRA and LRU approximations of the inverse Hessian, and thus the posterior covariance, is equal to the deficiency of the basis $\mathbf{Q}$ relative to the identity matrix. There will always be rank deficiency in $\mathbf{Q}$, so long as $l < n$. This is not a surprising outcome since the point of using the LRU is to prevent that exact issue.

The analysis increment is calculated from either the LRU or LRA by multiplying their respective inverse Hessian approximations by the right-hand-side of the least-squares problem, $\boldsymbol{b} \equiv \nabla_{\delta \boldsymbol{v}} J|_{\delta \boldsymbol{v}^k = \mathbf{0}}$. The difference between those increments is

$$\begin{aligned} \Delta_{\delta \boldsymbol{v}} &= \delta \boldsymbol{v}_{\text{LRA}} - \delta \boldsymbol{v}_{\text{LRU}} \\ &= \left(\mathbf{Q}\mathbf{Q}^\top - \mathbf{I}_n \right) \boldsymbol{b}. \end{aligned} \tag{E.4}$$

When the first basis vector is equal to the normalized gradient (i.e., $\hat{\mathbf{Q}} = \begin{bmatrix} \hat{\boldsymbol{b}} & \mathbf{Q} \end{bmatrix}$), the LRA-LRU increment difference simplifies to

$$\begin{aligned} \Delta_{\delta \boldsymbol{v}} &= \left(\hat{\mathbf{Q}}\hat{\mathbf{Q}}^\top - \mathbf{I}_n \right) \boldsymbol{b} \\ &= \hat{\mathbf{Q}}\hat{\mathbf{Q}}^\top \boldsymbol{b} - \boldsymbol{b}. \end{aligned} \tag{E.5}$$

Since

$$\hat{\mathbf{Q}}^\top \boldsymbol{b} = \begin{bmatrix} \|\boldsymbol{b}\| \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{l \times 1}, \tag{E.6}$$

and

$$\hat{\mathbf{Q}}\hat{\mathbf{Q}}^\top \boldsymbol{b} = \hat{\mathbf{Q}} \begin{bmatrix} \|\boldsymbol{b}\| \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{b}, \tag{E.7}$$

the increment difference further simplifies to

$$\Delta_{\delta \boldsymbol{v}} = \boldsymbol{b} - \boldsymbol{b} = \mathbf{0}. \tag{E.8}$$

Therefore, when $\hat{\boldsymbol{b}}$ is used as the leading basis vector, the difference between LRU and LRA analysis increments is always zero, and not only for this application.

Still, there is no particular reason why $\hat{\mathbf{Q}}$ would be a better (or worse) basis of the Hessian than $\mathbf{Q}$. We also have yet to demonstrate why an eigendecomposition produced with $\hat{\mathbf{Q}}$ would require fewer modes to converge on the increment produced with the full-rank Hessian than the truncated SVD.

Appendix F

Important processes in an NWP-Chemistry Model

NWP-Chem Processes:

- Advection and Diffusion
- Surface-air interactions (LSM)
- Turbulent Mixing (PBL)
- Cumulus Convection
- Emissions
- Wet and Dry Deposition
- Chemistry
- Aerosol Activation
- Aerosol Thermodynamics
- Radiation
- Microphysics