Data for Modelers

Helping Understand the Climate System

Mark A. Parsons
B.Sc. Cornell University 1988

A thesis submitted to the Faculty of the Graduate School of the University of Colorado in partial fulfillment of the requirement for the degree of Master of Arts in the Department of Geography, 2010.

This thesis entitled:
Data for Modelers—Helping Understand the Climate System
written by Mark A. Parsons
has been approved for the Department of Geography

_____
Barbara P. Buttenfield

_____
Mark C. Serreze

_____
Eric A. Kihn

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.

IRB protocol # 0909.36

Abstract

Parsons, Mark A. (M.A. Geography)
Data for Modelers—Helping Understand the Climate System
Thesis directed by Professor Barbara P. Buttenfield

The Arctic is changing rapidly with dramatic local and global effect. To understand that change requires understanding the Arctic as a system. Models of different processes and at various scales are necessary tools for analyzing and understanding the Arctic system. Models are extremely diverse, yet they all require quality data. Through a series of case studies, augmented with with ethnographic observation around the International Polar Year, this thesis examines how modelers assess, acquire, and prepare data for their models. By comparing specific case studies, common themes emerge that can be compared against broader observation. These themes, in turn, suggest data management techniques or requirements for data systems to improve access and use by modelers and generally improve understanding of the Arctic system. This case study based approach has proven to be a useful method for teasing out both general and specific data needs for different models. An overarching lesson is that greater short-term benefit to modelers and significant gains in efficiency can be achieved by improving the formats, convention, and consistency of the data rather than improved interfaces and analysis tools. A "data-first" philosophy can improve the data systems that support the overall interdisciplinary, integrative science necessary to understand the complex Arctic system.

Table of Contents

List of Tables

List of Figures

Chapter I
INTRODUCTION

The Arctic is undergoing dramatic climate change with significant impact on the people who live there (ACIA, 2005; Krupnik and Jolly, 2002). To understand this change and predict the future state of the Arctic, researchers need to take a synthetic and systemic approach that addresses the Arctic as a system. This integrative and interdisciplinary approach is a major focus of the NSF Arctic System Science program, the interagency Study of Environmental Change (SEARCH) (SEARCH, 2005), and the International Polar Year (Allison et al., 2007; ICSU, 2004). Significant advances in our understanding of the Arctic system have come through integrative studies. For example, Serreze et al. (2000) document pan-Arctic warming and geophysical changes such as retreating sea ice, snow, and glacier extent and permafrost thaw through a broad synthesis of observations spanning 400 years. Chapin et al. (2005) demonstrate how the increased prevalence of shrubs over Arctic tundra (Sturm et al., 2001) has reduced albedo and created a positive warming feedback. Overpeck et al. (2005) describe the "essential components" and interactions of the Arctic system and how a seasonally ice-free period in the Arctic could fundamentally alter that system.

Central in these and related interdisciplinary studies is the integration of data from multiple sources. Further advances will likely come from similar integrative studies that will rely on an array of models and other tools to integrate, interpret, and even augment disparate data. The importance of models was emphasized at a recent NSF-sponsored workshop: "Arctic System Synthesis Workshop: New Perspectives through Data Discovery and Modeling" (http://www.arcus.org/arcss/message_050707.html). In this context, the term "model" is a very broad term describing tools ranging from complex global circulation and numerical weather prediction models running on powerful centralized computers to targeted ecological distribution models run in a GIS on a researcher's laptop computer. Despite this disparity, all models have one thing in common. They require data.

Currently Arctic data are managed in disparate ways. Investigators spend undue time seeking data and preparing the data for analysis and there is a need for a more integrated approach to Arctic data management (ICSU, 2004; NRC, 2006; Parsons, 2006; Parsons et al., 2010;

SEARCH, 2005). It is necessary to establish a close and collaborative partnership between scientists, observing systems operators, data managers, archivists, and relevant research programs to ensure efficient preservation and effective use of Arctic data. The resultant system should be designed around the needs of data users and providers to ensure that it is simple, predictable, reliable, and readily extensible to address multiple disciplines and innovative use of the data.

Because of the importance of modeling in understanding the Arctic system, a partnership between data managers and modelers can develop initial requirements for a broader Arctic data system and begin to identify what data management techniques, integration methods, and activities serve the Arctic modeling community and contribute to interdisciplinary synthesis and improve predictions of the Arctic system. This study is an initial step toward building that partnership and toward identifying modeler needs.

This thesis presents a series of case studies detailing how specific modelers actually access and use data in particular scientific investigations. The case studies are augmented with broad ethnographic observation of the Arctic research community. The idea is that by comparing specific case studies, common themes may emerge that can be compared against broader observation. These themes, in turn, may suggest data management techniques or requirements for data systems to improve access and use by modelers and generally improve understanding of the Arctic system.

Chapter 2 provides the theoretical basis for this work and it's approach, and Chapter 3 describes the general methodology. Chapter 4 describes the case studies and lays out the primary evidence for the analysis in Chapter 5. Chapter 6 concludes with some discussion and initial conclusions on how data for modelers can be improved.

Chapter II
Background and Rationale

The diversity of models makes understanding their data requirements complex. It is necessary to establish a common framework for requirements development.

Anderson and Woessner (1992) describe a formal protocol for numerical modeling. Their direct application is specific to groundwater modeling, but their protocol can be broadly applied as a conceptual model for the development of models in general. Figure 2.1 presents a modified version of the steps in the protocol. Note the multiple stages in the process that require data. Also consider how different stages may require different forms of data from compiled information useful to prepare the conceptual model to a large volume of detailed arrays for numerical processing.

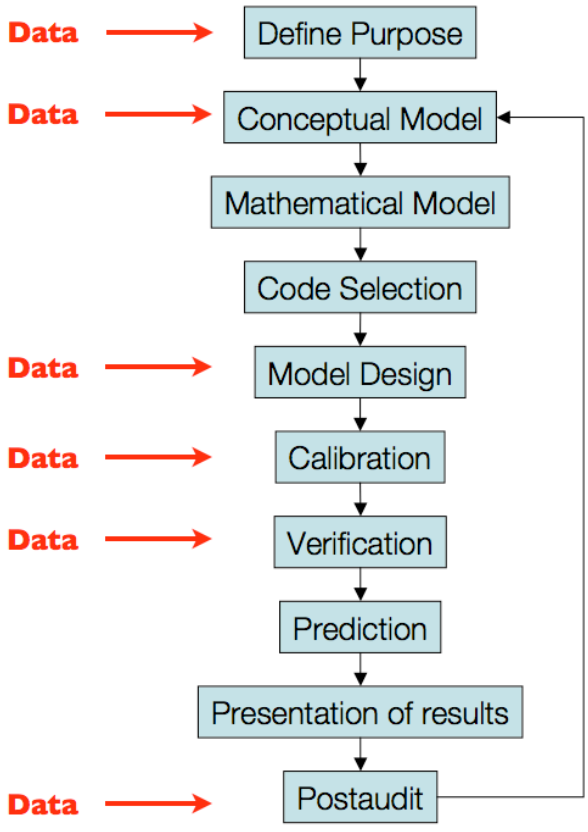

Figure 2.1. Simplified representation of the steps and data inputs in a protocol for model application. From Anderson and Woessner (1992).

At each of these stages requiring data, the modeler needs to go through the same basic steps.

1. Discover or identify the necessary data.
2. Assess the relevance, uncertainty, and quality of the data and their fitness for the application,
3. Acquire the data for processing and analysis. This could be a simple ftp transfer or physical acquisition of media or it could involve processing data remotely. In some applications real time or near real time acquisition is necessary.

4. Prepare the data for processing. This could involve digitizing analog records or reformatting, subsetting, gridding, interpolating, subsampling etc.

Data Discovery

Data discovery is an increasingly complex issue as the volume of Earth science data grows exponentially. Some of the challenges are technical, but many are rooted in the culture of individual geographic disciplines and the willingness of individuals to share their data (Key Perspectives Ltd, 2010; Parsons et al., 2010). The library, digital library, and data management communities have been researching this issue for a long time, and there are many national and international data discovery systems in place such as geodata.gov (formally the Geospatial One Stop) and the Global Change Master Directory (http://gcmd.gsfc.nasa.gov/). In addition, national and international Earth science organizations are increasingly pushing for greater data sharing and enhanced, standardized data descriptions or "metadata" to better enable discovery (de Sherbinin and Chen, 2005; ICSU, 2004; Nelson, 2009; OMB, 2002). Several projects and initiatives are explicitly addressing data discovery in the Arctic. These include the International Polar Year (Parsons and Wilson, 2007; Parsons et al., 2010), the Arctic Portal (http://www.arcticportal.com/), the Sustained Arctic Observing Network (SAON—http://arcticobserving.org), and the Cooperative Arctic Data and Information Service funded to support the Arctic Observing Network (NRC 2006). The other aspects of data handling for models (assessment, acquisition, preparation) have received much less attention.

Data Assessment

Uncertainty and error are inherent in geographic information not only because of limitations in data collection or analysis, but also because of imperfect human knowledge (Couclelis, 2003). While it is important to continually strive to reduce uncertainty, we must also recognize that there are limits to what we can achieve, especially in an environment of expanding data use. The ISO standard Open Archive Information System Reference Model requires that archives ensure their data are independently understandable by a designated user community (ISO, 2003). Yet user communities for a given data set can change over time, and some applications (including modeling) may be inappropriate for that data set. It is necessary for data managers to provide the necessary context for users to understand the limitations and appropriate use of data (Parsons and

Duerr, 2005). Providing and enhancing this context is an important aspect of data stewardship (NRC, 2007).

Understanding data quality was a major theme at the NSF Arctic Synthesis workshop mentioned earlier. Participants emphasized the need to be able to consult experts on the data and also suggested a variety of data "peer-review" schemes. The need to consult experts has been a recurring theme in recent data system development efforts (e.g., NRC 2007; Parsons and Wilson 2007). Peer-review of data is also a growing topic in the Earth science data management community (Parsons et al., 2010). A new journal, *Earth System Science Data*, has even been established as a means to publish high-quality data and all its relevant documentation in a classically peer-reviewed form.

In addition to data quality considerations, modelers have additional assessment criteria, such as whether the data are at an appropriate scale, have the necessary temporal and spatial coverage, etc. It is necessary for data managers and providers to understand how modelers assess data in order to provide the necessary supportive information, tools, and context in a meaningful way.

Data Acquisition

Data acquisition can be a relatively straight forward, but it is also hindered by many of the technical and cultural barriers that restrict data discovery. If data cannot be discovered then clearly they cannot be acquired, but sometimes a data description may be found but the data themselves are unavailable. Data may be inaccessible because of legitimate concerns about human privacy or threats to species, because they are not in a readily usable form (i.e. not digital), because they are not be available soon enough because of data provider imposed restrictions, or simply because the data were lost. For example, in 1998, the International Permafrost Association and World Data Center (WDC) for Glaciology, Boulder compiled a collection of metadata describing frozen ground related data housed around the world. In 2003, the WDC attempted to contact the investigators and institutions holding the 89 products not housed at the WDC. Forty-five of the 89 products were not readily accessible and may no longer be available (Barry and Smith, 2004). Data acquisition is also closely linked to the final step of

data preparation. When data can be readily manipulated remotely, transfer loads and data preparation requirements can decline.

Data Preparation

Data preparation may be the most time consuming step of the four. Because of the diversity or lack of Arctic data, modelers must address fundamental issues of scale and coverage and detailed technical issues of data formats and grid specifications to ensure data can be used in their model. Some of these issues are well understood and can be addressed through tools and techniques such as automated subsetting and reformatting, but project or discipline-specific models can use highly specialized data structures, resolutions, and time/space domains (e.g. hunting tags vs. polar orbiting satellite data). Furthermore, data preparation requirements will vary depending on where in the modeling protocol data are being used. Data managers need to be able to determine effective means to integrate data across space and time and facilitate ready data use by diverse modeling communities.

A Need for Best Practices

It is important to note that given the huge disparity of models in scale, discipline, and application, it is unlikely that any one data management approach or technical solution will solve the needs of Arctic modelers. So the question becomes whether it is possible to identify common themes and best practices to guide the development of existing and future Arctic data management systems as a whole. By comparing the needs of several different models, it may be possible to identify some of those themes and best practices.

Chapter III
OBJECTIVES AND APPROACH

The overarching goal of this thesis is to determine what data management practices, methods, or techniques assist Arctic system modelers. The central proposition is that there are common and instructive themes in how modelers assess, acquire, and prepare data. Examination of these themes can reveal first principles or overarching guidelines for Arctic data management. These principles, in turn inform data management practice in specific ways to improve Arctic system modeling. This study examines three disparate modelers/models exploring different phenomena at different scales to determine common themes and first principles. The analysis and results should be informative to any Arctic data manager, but they are also geared specifically to the National Snow and Ice Data Center where appropriate.

Model Comparison

A central issue in designing a data system is understanding how a scientist will need to use the system and for what purpose and then designing the system to meet those needs. Assessing user needs is challenging when trying to design a system that will provide a broad array of disparate data to users with differing expertise. Arctic research provides a particular challenge with its emphasis on interdisciplinary research in the physical, life, and social sciences. Comparing and contrasting the needs of different modelers can provide initial information that can be important first step toward the development of a broader interdisciplinary data system. A key question, then, is which models to compare.

One could take a targeted approach and examine several similar models in detail. For example, assessing several mesoscale snow models could lead to the development of fairly specific requirements and data needs for snow and potentially other land surface modelers. It would be difficult, however, to determine which of these needs are specific to the particular issue of modeling snow and which are broader issues that apply to Arctic modeling and synthesis in general. For example, addressing the spatial heterogeneity of snow cover or water equivalent may be such an overriding concern that it could overshadow other needs such as temporal consistency. Similarly, inputs to related models are likely to be in similar formats and have

similar data preparation issues thereby missing what may be fundamental data preparation issues for other models.

Another approach would be to compare related models that are used in conjunction to address a particular science question. For example a sea ice model and polar bear migration model might be used in conjunction to understand the effect of declining sea ice on polar bear populations (Durner et al., 2009). This would be an interesting study of how models can interrelate and provide input to each other. The issue of relating the models, however, could become the dominant question and obscure the more generic issues of how modelers assess, acquire and prepare data. Therefore, the approach here is to examine several disparate models in an attempt to identify common themes and best practices that reach across disciplines, scales, and modeling approaches.

Case Studies and Their Analysis

Various techniques can be employed to gather user needs, but not all are suitable in this situation. Surveys can provide a broad perspective but are limited in the depth of their analysis and cannot readily respond to issues identified by participants but unforeseen by the investigator. An historical or archival analysis can reveal user trends and preferences, but the disparate nature of Arctic data and how they are managed prevents a consistent analysis. Case studies, on the other hand, explicitly consider the context of a situation and provide a flexible model that allows investigators to probe more deeply into unanticipated areas of interest (Yin, 2003).

Yin (2003) notes three kinds of case studies: exploratory, descriptive, and explanatory. This investigation, while exploratory, also seeks to be both descriptive (how do modelers work?) and explanatory (what practices could help modelers?). Accordingly, this study takes a multi-facetted analysis strategy as well. In Chapter 4 each case is presented and examined in a consistent descriptive framework. The framework considers the process each modeler underwent in the context of the model presented by Anderson and Woessner (1992) and carefully considers data assessment, acquisition, and improvement at each stage of data use in the process. This description provides a general understanding of how the modeler works. In Chapter 5, each case is analyzed individually once immediately after it is conducted and then again after each

subsequent case study is conducted. The individual assessments identify possible areas for data system improvement in each study. The last section of Chapter 5 presents a cross case comparison and synthesis. The comparison identifies common issues and themes across the different models. The synthesis then examines the areas for improvement and common themes to develop and test more refined propositions.

Affirming the propositions is a matter of finding multiple lines of evidence that support it. This document provides the overall summary of the evidence, but each case study has multiple sources of evidence. Sources of evidence include notes, recordings and follow-up on the interviews, the published papers of each modeler, other literature, and the data and interfaces used by the modelers (Appendix C provides additional information on each of the data sets). Finally, to test external validity, I further tested each proposition against experience over the last five years helping lead an effort to define an international, interdisciplinary data system for the International Polar Year and beyond (de Bruin et al., 2009; Chen and Parsons, 2010; LeDrew et al., 2008; Parsons, 2006; Parsons et al., 2010). The question was simply did the proposition contradict or was it supported by any of the conclusions growing out of the experience of the dozens of data managers involved in the International Polar Year Data and Information Service. Although IPY was much broader than modeling, its very interdisciplinary focus was likely to face similar issues. It is also reasonable to assume that the IPY experience influenced the formulation of my theories and assertions.

Detailed Approach

The specific approach is as follows.

1. Identify a pan-Arctic scale sea-ice and climate model, a meso-scale land surface model, and a plot-scale ecological resource model with modelers willing to participate. Chosen models are described in chapter 4.

2. Develop the common case study design across the multiple case studies built around the key study questions of how modelers assess, acquire, and prepare data at the various stages in the modeling process. While later case studies may change slightly in response to discoveries in an earlier study, it is important to maintain some consistency to ensure

external validity and reliability of the results (Gerring, 2007; Yin, 2003). Appendix B includes the case study survey protocol. Research design elements include:

a. An assessment of whether the modeler followed a consistent process such as described by Anderson and Woessner (2005), a description of the process followed, and the points in the process where the data are required.

b. A common set of interview questions focused on understanding the overall process including core topics, such as

    i. How they define and assess data quality.

    ii. To what degree they rely on the original data creator or other experts for guidance or authority on data.

    iii. Attributes (e.g. scale, parameters, time step) necessary to assess applicability of data to a model.

    iv. Use of interfaces and tools to assess and access data.

    v. Their own use of software tools, platforms, and expertise.

    vi. Specific data requirements of their model (e.g., data models or formats, grids, interpolations, projections).

    vii. Downscaling and/or upscaling techniques.

    viii. Calibration and validation approaches

c. Documentation of the science question addressed by the modeler in their work either historically or currently.

d. A site visit and interview.

e. Follow up phone and e-mail interviews with each modeler.

3. Describe the three case studies in a consistent framework including textual descriptions and diagrams for each mode describing the study process, data use, and how the data were assessed, acquired, and prepared for use. Review with case study participants.

4. Analyze each case for specific data system needs or requirements. Repeat after each study is conducted.
5. Analyze all three cases, seek patterns, and identify common themes.
6. Formulate and test first principles, best practices, or techniques against multiple sources of evidence: the interview, participant publications, other literature, and the data used.
7. Test against ethnographic observation and the experience of the International Polar Year.
8. Present the results at relevant conferences and other fora such as AAG, the Arctic Research Consortium of the U.S., and the Earth Science Informatics section of AGU, and seek feedback from the community.
9. Publish the results in the literature and other outlets for the scientific and data management communities.

Subsequent chapters describe how this approach was implemented. Chapter 4 introduces the models and describes each case in a consistent framework. Chapter 5 provides the analysis and exploration of propositions and principles. Chapter 6 concludes with some discussion and initial conclusions on how data and systems for modelers can be improved.

Chapter IV
The Models and Case Studies

Three models form the basis for the case studies: The *Community Sea Ice Model version 5* (CSIM) (Briegleb et al., 2004), which is the sea ice component of the global-scale Community Climate System Model version 3 (CCSM); *SnowModel* (Liston and Elder, 2006a), which aggregates several local-to-regional-scale submodels to simulate multiple snow processes; and the *Multiple Element Model* (MEL) (Rastetter and Shaver, 1992; Rastetter et al., 1997), which compares plot-scale interactions between the cycles of two ecosystem elemental resources (e.g., carbon and nitrogen). Table 4.1 summarizes the spatial scale, predictive variables or science domain, and the general application of each model. As discussed in Chapter 3, the intention was to have very disparate models and applications. The models range in spatial scale from meters to 100s of kilometers and can each be considered at a variety of time scales. In addition, models from both the life and physical sciences were chosen to enable exploration of different disciplinary cultures and approaches to data handling and processing. The models also vary in their general types of application. For example, Snow Model is typically used in a very applied context to get the best possible representation of spatially variable snow properties across an area, while MEL is generally used in a more theoretical context to understand biogeochemical processes in ecosystems.

Table 4.1 Overview of the three models that serve as the basis for the case studies.

Model	Spatial Scale	Predictive variables	General application
Community Sea Ice Model (within the Community Climate System Model)	1 - 3°	• ice thickness distribution • ice area • ice volume • ice internal energy • snow volume • surface temperature • ice velocity • stress tensor components	Prediction of different potential future sea ice regimes based on different forcing scenarios.
SnowModel	1 m - 10 km	• snow accumulation • blowing-snow redistribution and sublimation • forest canopy interception, unloading, and sublimation • snow-density evolution • snowpack melt	Detailed characterization of snow properties distributed across an area based meteorological, topographic, and vegetative parameters.
Multiple Element Model	defined plot	• stocks and fluxes of two elements (i.e. N and C) within the plot.	Improved theoretical understanding of biogeochemical ecosystem processes.

A specific model application and science question was chosen to form the basis for each case study. Each of the modelers— Marika Holland, National Center for Atmospheric Research; Glen Liston, Colorado State University; and Edward Rastetter, Marine Biological Laboratory—agreed to participate in the study and helped identify the relevant application and science question. Subsequent sections in this chapter provide an overview of each specific model application; a description of the process each investigator went through; a description of the various data used in the study; and discussion of how each investigator assessed, acquired, and prepared their data for use.

The Community Sea Ice Model

Overview

The Community Sea Ice Model (CSIM) is the sea ice component of the Community Climate System Model version 3 (CCSM), which also includes atmosphere, ocean, and land surface components (Collins et al., 2006). CSIM v5 captures five state variables across a five-category ice thickness distribution: sea ice area, sea ice volume, sea ice internal energy, snow volume, and

surface temperature. In addition, it captures ice velocity and stress tensor components, but these are not resolved across the ice thickness distribution. Boundary conditions are generally represented as ocean and atmospheric fluxes and state variables from the coupler to the other components of the CCSM. Alternatively, the CSIM can provide boundary conditions to other CCSM elements (Briegleb et al., 2004). This means that data requirements are dependent upon the scenarios being modeled. This case study examines how Holland et al. (2006) compared multiple ensemble runs of the CCCSM and other models to predict future reductions in Arctic sea ice.

Dr. Holland's paper is sometimes referred to by the Arctic research community as the sea-ice "tipping point" paper. A tipping point is a transition from one stable climatic state to another. Dr. Holland says they began the study by looking for a tipping-point, partially in response to Lindsay and Zhang's (2005) initial investigation that asked "The Thinning of Arctic Sea Ice, 1988–2003: Have We Passed a Tipping Point?". Dr. Holland and her team began an initial examination sea ice time series from different runs of the coupled CSIM. One run, which they highlight as the first figure in their paper (Figure 4.1.), stood out as a clear example of rapid ice loss. This prompted more detailed analysis not only of the different CCSM ensemble members but also 15 other models archived as part of the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset.

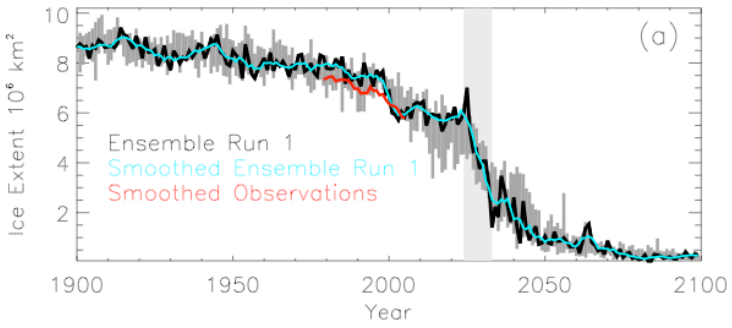

Figure 4.1. "Northern Hemisphere September ice extent for one Run 1 (black), the Run 1 five-year running mean (blue), and the observed five-year running mean (red). The range from the ensemble members is in dark grey. Light grey indicates the abrupt event." (Holland et al. (2006), p.2) ©2006 by the American Geophysical Union. Used with permission.

Figure 4.2 outlines the general process for Dr. Holland's study and indicates what data are used at different points in the process. Table 4.2 provides more specifics on the data. The steps in Figure 4.2 are as follows.

0. The first step, of course, was to run the model, but in this case running the model was not an explicit part of the study. The model runs used were production runs that would have been run for the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment or other purposes, regardless of Dr. Holland's study (hence step "0").

1. The first actual step in this study was to identify and acquire the relevant model outputs. Dr. Holland considers the outputs from the model runs as the first data input into the process.

2. Create single variable time series. The CCSM typically produces a single monthly file for all variables, so it was necessary to produce time series of relevant variables such as sea ice area or thickness.

3. Conduct initial assessment. Based on those time series, investigators do an initial, largely qualitative, assessment comparing the model outputs to observations.

4. Conduct quantitative assessment. Based on the initial assessment, investigators conducted more rigorous quantitative comparisons between model output and observations. Note that different observations are used in the quantitative analysis than in the qualitative analysis. The investigators have a standard diagnostic package that includes multiple variables but the only *observed* value used in the package is the 15% ice concentration line (as a measure of ice extent).

5. Conduct statistical analysis. Building off the quantitative comparison with the observations, investigators conduct detailed statistical analyses of the modeled time series to try and determine the key driving mechanisms common across model runs. This analysis may lead to re-running the model with different simulated atmospheric conditions or forcings. Only the CCSM is run repeatedly not the other CMIP3 models.

6. Present results in papers, talks, and case study interviews.

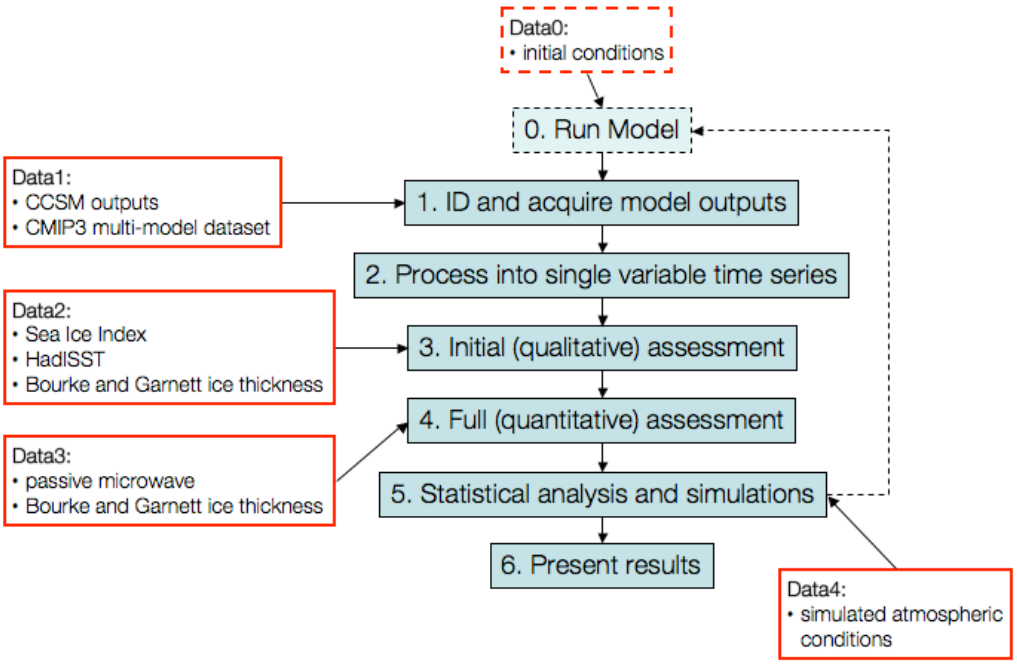

Figure 4.2. General workflow and data inputs for the research process leading to Holland et al. (2006).

Data enter this process at several stages. Initial input to force the models (Data0) are predefined forcings and boundary conditions used in the CCSM and are not considered in this study. Data1 are the model outputs central to the investigation. Data2 and Data3 are generally observations used to compare against the models. Data4 are predefined simulated future atmospheric conditions (IPCC, 2000).

Table 4.2 provides details on the particular data collections. The data are disparate and include satellite and submarine observations, climate and chemical transport model outputs, reconstructions combining models and observations, and published predicted scenarios of climate change. By their nature, observations vary more than the model outputs in scale and coverage, but they also less likely to adhere to a common set of data standards. The climate model outputs are very well defined either through formal model coupling (e.g., CSIM is directly coupled to the atmosphere, ocean, and land components of the CCSM) or because of the large effort conducted by the WCRP CMIP3 to define and collect common data formats, grids, and variable names for climate models. In this context, observations require greater assessment of

usability and more effort to acquire and prepare for comparison with the models. Therefore Data2 and Data3 receive more attention in subsequent discussion.

Table 4.2 Data collections used in Holland et al. (2006) as shown in Figure 4.2. See extended table in Appendix C.

Stage	Data Collection/Data Set	Application	Data Source
Data0	inputs to CMIP3/CCSM models	forcing, parameterization	n/a (production runs done by others)
Data1	CCSM ensemble runs	analysis	local
Data1	WCRP CMIP3 multi-model dataset (15 models)	analysis	PCMDI/Earth System Grid: https://esg.llnl.gov:8443/
Data2	Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST) (Rayner et al., 2003)	initial assessment	NCAR Research Data Archive http://dss.ucar.edu/datasets/ds277.3/
Data2	Sea Ice Index (Fetterer and Knowles, 2002)	initial assessment	NSIDC: http://nsidc.org/data/g02135.html
Data2	Ice thickness (Bourke and Garrett, 1987)	initial assessment	Bourke and Garett, 1987
Data3	Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I Passive Microwave Data (Cavalieri et al., 1996)	full assessment	NSIDC: http://nsidc.org/data/nsidc-0051.html
Data4	Special Report on Emissions Scenarios forcings	simulation forcing	standard simulations (IPCC, 2000)

The WCRP CMIP3 Multi-Model Data Set Archive

Before examining how Dr. Holland conducted her data assessment, some discussion of the CMIP3 Multi-Model Data Set Archive is in order. As mentioned, the WCRP made a substantial effort to harmonize data across all the CMIP3 models. This effort has significant impact on this and other studies by Dr. Holland and others. It is, therefore, helpful to have some understanding of the CMIP3 model archive when examining Dr. Holland's work.

The Program for Climate Model Diagnosis and Intercomparison (PCMDI) at the Lawrence Livermore National Laboratory (LLNL) collects model output contributed by modeling centers around the world as part of an effort organized by the WCRP's Working Group on Coupled Modelling (WGCM). This effort supported CMIP3 and was intended to serve scientists preparing the Fourth Assessment Report of the IPCC. In particular, it is meant to support IPCC's Working Group 1, which focuses on the physical climate system. The collection includes outputs that are simulating the present climate, the historical climate of the 20th century, and future climates that may occur in response to various scenarios of future greenhouse gas emissions. It is formally referred to as WCRP's CMIP3 multi-model dataset and is supported by

the Office of Science, U.S. Department of Energy. More details are available at http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php.

There are very detailed requirements for the format and representation for the model outputs that are submitted to the archive. Data must be rpresented in a regular longitude-latitude Cartesian grid in the netCDF-CF format. Each file must contain only a single output field from a single simulation (i.e., a single run). There is a standard set of variables, which must be named according to a defined convention, and the coordinate variables must use particular units in a defined way. For example, latitude and longitude must be expressed in "degrees_north" and "degrees_east" respectively. Detailed metadata is also required. This required remarkable community effort. There was little direct incentive to make this effort beyond the motivation to contribute to the IPCC and to enable broad use and scrutiny by the international scientific community thereby improving the models over the long term. In the observational community, there is apparently less willingness to do this same level of harmonization for observations, and indeed it would be a much greater effort. In the Arctic Observing Network (AON), for example, most data are being submitted to the data system in very diverse, custom ASCII and even proprietary (e.g., Microsoft Excel) formats. The data system supporting AON encourages the use of netCDF to better enable data integration, comparison, and visualization but so far the AON investigators have shown little interest. As we will see in subsequent analysis, this disparity between heterogeneous observational data and well-controlled model inputs and outputs can be a key issue.

Data Assessment

Data1: Assessment for Data1 is a scientific assessment. There is not really a need to choose between different data sets, it is simply a matter of using what is available locally through the CCSM and through what is available in the CMIP3 archive led by the Program for Climate Model Diagnosis and Intercomparison (PCMDI). The PCMDI archive is a remarkable resource for the research community. Dr. Holland said it "has changed the way people do their studies". A similar level of coordination in Arctic observations would be very helpful.

Data2 and Data3: These data are used to assess the performance of a given run against the observational record. There are several criteria to assess which observational data would be most appropriate as shown in Figure 4.3. Of course, the first criterion is the geophysical parameter being examined. This is constrained by what the models produce, but is also limited by what observations are available. This study primarily looked at ice extent. They included some consideration of ice thickness but were limited by the availability of suitable observations. Spatial and temporal coverage requirements (whole Arctic and as long and current as possible) further limit the options to few choices. Then in the initial assessment stage the convenience of common data formats and grids drive one set of choices, while for a the full assessment it becomes necessary to use the most accurate available product, even though it will require additional work to reformat and interpolate the data on another grid.

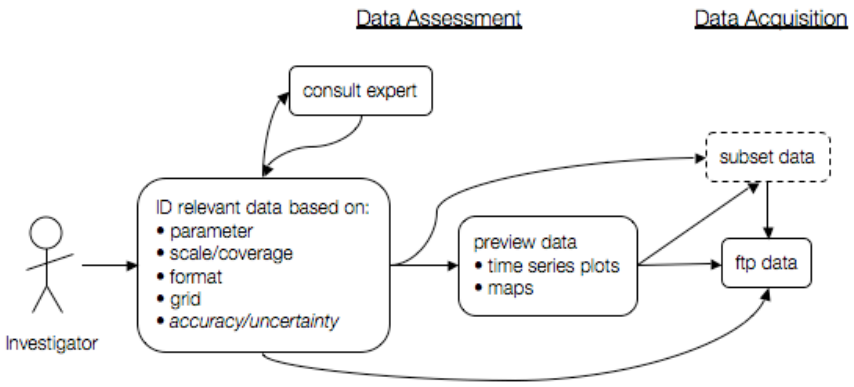

Figure 4.3. Basic data assessment and acquisition process for Data2 and Data3 in Holland et al. (2006).

The HadISST product is readily comparable to the model outputs because it is in the same precisely defined, self-describing format (netCDF with Climate Forecast extensions), and Dr. Holland can quickly view it with tools she has to hand (e.g., ncview), but Dr. Holland's colleagues in the sea ice research community advised her that the passive microwave time series was more consistently accurate for spatially broad, time-series comparisons. As a result, HadISST was used in the initial assessment, but the passive microwave data was used for the quantitative comparison shown in Figure 4.2. Interestingly, the Sea Ice Index which is an interpretive product consisting of images, maps, and time-series plots is derived from the passive microwave data. This enabled quick qualitative comparison, but when it came to quantitative comparison, the raw data were downloaded, reformatted, and regridded.

The Sea Ice Index also provided another important function by providing an inherent pointer to a specific sea ice data set. The National Snow and Ice Data Center provides dozens of sea ice products, including several products derived from passive microwave brightness temperatures that differ subtly by using different algorithms and error corrections (http://nsidc.org/data/seaice/). Dr. Holland, while aware of the differences, is not well versed in the subtleties. She was not sure which of the two primary time series she used, only that she got what was available through the Sea Ice Index. As Dr. Holland said, "I have enough data to look at," and "My work is not to compare data sets". The point is that she typically wants a defined benchmark data set, often a climatology, readily available in netCDF-CF in a 1° latitude/longitude grid. She relies on colleagues and scientific experts (not data centers or services) to advise her on the most appropriate product and only reformats and regrids when she must. While, we focused on sea ice concentrations and extent in the interview, the same approach generally holds for other parameters (e.g., ice thickness or velocity) and forcings (e.g. radiative fluxes).

Data Acquisition

In this study, analysis was done locally on individual workstations and data were stored on the NCAR mass storage system. The organized, centralized facilities of NCAR encourage investigators to directly acquire the data. For example, all data except quick looks were transferred via ftp even though some data such as the CMIP3 data set are available through the Open-source Project for a Network Data Access Protocol (OpenDAP) and are directly accessible remotely through IDL, a primary research tool of the investigators. The sea ice observations are not currently available through OpenDAP, but it would be worth exploring that possibility as a mechanism to address the data format requirements of the modeling community.

Data Preparation

Data preparation consisted of three possible processes: 1) creating single-variable time series from gridded fields, 2) interpolating and regridding data to the CSIM grid, and 3) reformatting data to netCDF-CF. Creating the time series was only necessary for the CCSM output and is a routine part of Dr. Holland's job assessing CSIM output. Interpolating and regridding, while an

effort, is also fairly routine, because the CSIM uses an unusual grid with the North Pole displaced into Greenland to reduce the convergence of meridians in the Arctic Ocean—the area of study. Reformatting, however, is a much bigger issue. Just recently, Dr. Holland, tried to do a comparison with some of the SSM/I data and some of her scripts didn't work because of troubles with the format. This can create a level of frustration that can delay or even prohibit initial comparisons that could bear fruit.

Summary

Dr. Holland examined multiple runs of the CCSM and other models from the CMIP3 looking for abrupt shifts in summer sea ice and comparing them to the observational record. The model data she examined is stored at the WCRP CMIP3 Multi-Model Data Set Archive, an invaluable resource for climate modeling studies. The CMIP3 effort has done much to standardize formats, grids, and conventions for climate model output. When comparing the model outputs to observations, Dr. Holland conducted a two-part assessment process, an initial qualitative assessment followed by a more rigorous quantitative assessment as appropriate. This two-step process was necessary because the more consistently accurate time series, the NSIDC sea ice concentrations, was in a less convenient format and grid. So it was easier to do an initial qualitative assessment with the HadISST data because they were in the same format as the model output. Investigators conducted their analysis on local workstations, but they also had the advantage of direct access to the large NCAR mass storage system. Dr. Holland's primary tools were IDL and various Unix shell scripts.

SnowModel

Overview

SnowModel aggregates three submodels: a surface-energy balance model, a snowpack evolution model, and a wind driven snow depth evolution model. It simulates multiple snow processes include snow accumulation, redistribution, sublimation, density evolution, and snowpack melt for the global snow classes defined by Sturm (1995). The model was designed to be applicable in many different landscapes and has been applied in Greenland (Mernild et al.,

2006), Antarctica (Liston and Winther, 2005), and forested landscapes (Liston and Elder, 2006a; Liston et al., 2008a). It requires as inputs meteorological time series and spatially distributed vegetation and topography. The reliability, availability, and consistency of these data can be problematic in the Arctic (NRC, 2006).

This case study examines how Liston et al. (2008a) used SnowModel to simulate snow accumulation and distribution across three study areas with extensive observations made during the NASA/NOAA Cold Land Processes Experiment (CLPX). CLPX was a field study conducted in Northern Colorado and Southern Wyoming between fall 2002 and spring 2003. The experiment was designed to improve quantitative understanding, models, and measurements necessary to extend our local-scale understanding of cold-region water fluxes, storage, and transformations to regional and global scales. It explored the relationships between process-oriented understanding, land-surface models, and microwave remote sensing by using a multi-sensor, multi-scale approach. Intensive ground, airborne, and spaceborne observations were collected within a framework of nested study areas ranging from 1 ha to 160,000 km2. Study areas were selected to represent diverse snow regimes, including that of the Arctic tundra. Four Intensive Observation Periods (IOPs) were conducted during February and March of 2002 and 2003 (Cline et al., 2003).

The authors of Liston et al. (2008) were some of the leaders of CLPX and were largely responsible for the experiment design and implementation. Glen Liston was the general modeling lead for the experiment, Kelly Elder was in charge of the field data collection, and Donald Cline was the experiment's principal investigator and leader of the airborne remote sensing data collection. The Liston et al. (2008) study was one culminating result of the CLPX, in that it produced the best possible representation of high-resolution spatial (30 m) and temporal (daily) distribution of snow water equivalent (SWE) over the 25 x 25 km CLPX Mesoscale Study Areas (MSAs). A central goal of CLPX was to produce a legacy data collection and the daily, 30-m MSA grids produced by Liston et al. (2008) are a significant contribution to that collection.

Dr. Liston said the study was also the first complete application of a unified modeling system that incorporates the three submodels and a meteorological distribution model, MicroMet (Liston

and Elder, 2006b), in a standard and readily configurable way. Dr. Liston emphasized that the modeling system is now a coherent package with one menu that assumes standard inputs and is flexible enough to apply to any piece of cold-region real estate regardless of terrain, vegetation, snow type, or scale (Dr. Liston has other studies in progress applying SnowModel at scales that range from 10 cm to 10 km grid increments). Over the years, Dr. Liston had been developing the various components of the system and modifying them to handle different environments. For example, SnowModel had originally been applied only in unforested, high-latitude environments, but for CLPX, Dr. Liston improved the model to address important forest processes such as vegetation snow catch and sublimation (Liston and Elder, 2006a).

Producing this flexible, unified, modeling system was a major objective for Dr. Liston under CLPX, and in some ways it was driven by the requirements of creating this legacy data set. In order to create the best possible data set, Dr. Liston tried to gather as much relevant input data as he could. Having all this data in many different formats drove him to develop "a tool that was smart enough to take any given meteorological tower or any given data set and run my analysis scripts and QC procedures." The final study, therefore, involved a complex interplay of models, assimilation schemes, QC processes, and input data sets.

Figure 4.4 outlines the general process for Dr. Liston's study and indicates what data are used at different steps in the process. Table 4.3 provides more specifics on the data. The process is not really as linear as shown in Figure 4.4, and some minor steps are missing, but this provides a basic overview in a similar manner as used in the Holland case study.

1. The process begins by identifying data available from a myriad of meteorological stations within the three CLPX Meso-scale Study Areas. Dr. Liston identified 27 relevant stations. Some were installed as part of CLPX; others were part of other monitoring networks. The stations are summarized in Table 4.3 and are described in more detail in Liston et al. (2008, Figure 1 and Table 5).
2. Determine which variables to use from each meteorological station. MicroMet requires inputs of air temperature, relative humidity, wind speed, wind direction, and precipitation, but Dr. Liston did not use all these variables from all stations. For example,

temperature was used from all the stations, while precipitation was not used from any of them. Precipitation forcing came from other sources. Which variables were used from each station is described in Liston et al. (2008, Table 5).

3. Process the meteorological data into a common hourly format for input into MicroMet. Each data source required its own set of processing scripts.

4. Correct all meteorological data for missing values and out-of-range or spurious values in accordance with standard protocols.

5. Run MicroMet to distribute all the meteorological variables over space and time and to provide daily, gridded values for input into SnowModel

6. Run SnowModel (including all its subcomponents) to produce initial estimates of SWE values and distribution.

7. Prepared observed SWE data collected as part CLPX by producing areal averages and by applying corrections to some of the data. This step could occur at any time earlier in the process, but it is necessary for the next data assimilation step.

8. Run SnowAssim, a methodology for assimilating observed snow data within SnowModel (Liston and Hiemstra, 2008). SnowAssim could be viewed as a subcomponent of SnowModel that forces the modeled results to match up with observed values when and where they occur.

9. Create a spatially distributed, precipitation-correction factor, based on the differences between the modeled results and the observed values by fitting a surface across the differenced values and their locations. These surfaces are shown in Liston et al. (2008, Figures 5b, 6b, and 7b). Apply this correction factor to the LAPS precipitation values.

10. Run the entire modeling system of MicroMet and SnowModel again with the corrected precipitation values. In this study it was only necessary to apply the precipitation correction once, but sometimes, especially in blowing snow conditions, it can be necessary to recalculate and apply the precipitation correction.

11. Present the final spatial and temporal distribution of SWE in papers, talks, and case study interviews.

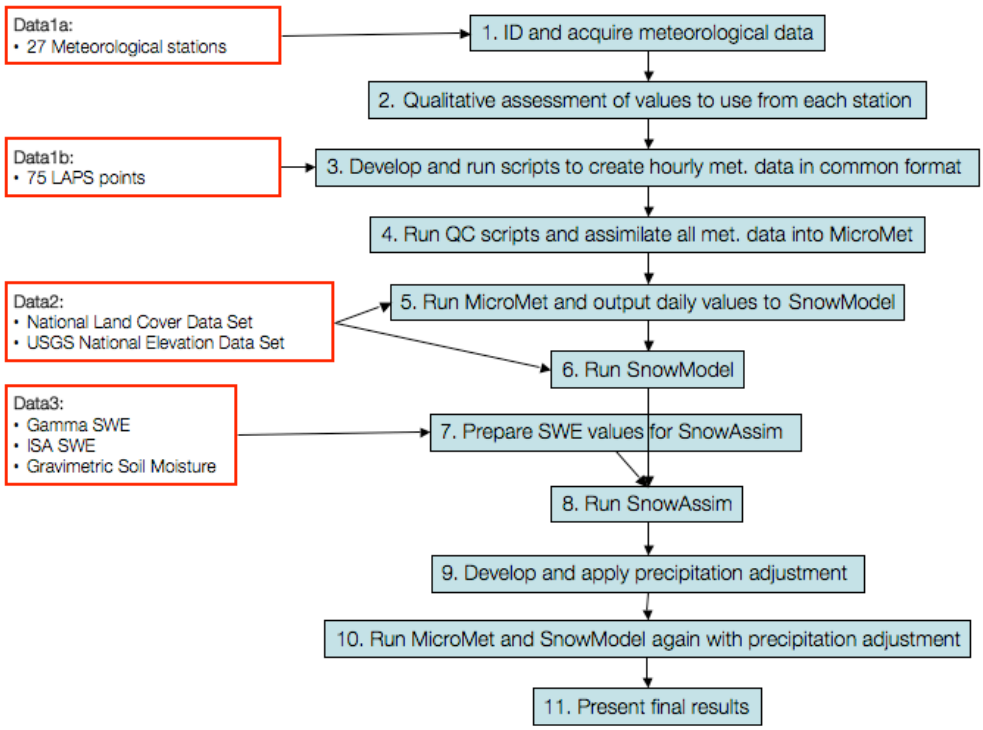

Figure 4.4. General workflow and data inputs for the research process leading to Liston et al. (2008).

Table 4.3 provides details on the observational data and model output used in the study. Data1 consists of five basic meteorological inputs—air temperature, relative humidity, wind speed, wind direction, and precipitation—into MicroMet, which aggregates the hourly, point-based inputs into daily values and distributes them across a regular 30-m grid. Data1a are observations from 27 meteorological stations from six different networks or organizations. Not all variables were used from every station. Liston et al. (2008, Table 5) provide details.

Data1b are gridded atmospheric analyses from the Local Analysis and Prediction System (LAPS (Liston, 2004; Liston et al., 2008b)) run by the NOAA's Earth System Research Laboratory. LAPS combines numerous observed meteorological data sets into a collection of atmospheric analyses. The observed data inputs into LAPS are not considered here, because

LAPS was run over the CLPX Large Regional Study Area separately as part of CLPX and is not directly part of Liston et al. (2008). LAPS analyses include many different 2-D and 3-D variables, but Dr. Liston only used the five surface field variables required by MicroMet.

Data2 are standard, reference topographic and vegetation-classification data sets used by both MicroMet and SnowModel. Data3 are observational data collected as part of CLPX. One data set was SWE calculated from an average of many field measurements collected during two Intensive Observation Periods at nine 1 km^2 Intensive Study Areas (ISAs)—three within each MSA. Another data set is SWE derived from a standard airborne remote sensing technique that measures Gamma radiation coming from the Earth. This data set had a correction applied using the gravimetric soil moisture data collected in one of the MSAs.

Table 4.3 Data collections used in Liston et al. (2008) as shown in Figure 4.4. See extended table in Appendix C.

Stage	Data Collection/Data Set	Application	Data Source
Data1a	Met. data from 10 main CLPX Stations	MicroMet forcing and assimilation	As described in Elder et al. (2009) based on Elder and Goodbody (2004).
Data1a	Met data from 5 Fraser Experimental Forest Stations Data	MicroMet forcing and assimilation	Personally from K. Elder
Data1a	Met. Data from 9 National Resource Conservation Service Snow Telemetry (SNOTEL) Stations	MicroMet forcing and assimilation	internet: http://www.wcc.nrcs.usda.gov/snow/snotel-temp-data.html
Data1a	Met data from the CLPX flux tower	MicroMet forcing and assimilation	internet
Data1a	Met data from the National Resource Conservation Service Dry Lake Remote Automated Weather Station (RAWS)	MicroMet forcing and assimilation	internet
Data1a	Met. Data from the Desert Research Institute Storm Peak Station	MicroMet forcing and assimilation	internet
Data1b	Local Analysis and Prediction System (LAPS) analyses	MicroMet forcing and assimilation	Locally as described in Liston et al. (2008). Also available at NSIDC (Liston, 2004)
Data2	USGS National Elevation Dataset	MicroMet and SnowModel forcing	http://ned.usgs.gov/
Data2	USGS National Land Cover Database	MicroMet and SnowModel forcing	Vogelmann et al. 2001; http://www.mrlc.gov/
Data3	Average SWE from ground measurements over the CLPX Intensive Study Areas (ISAs)	Snow Model assimilation	Personally from K. Elder as described in Elder et al. (2008). Calculated from data held at NSIDC (Cline, et al., 2003a; Cline, et al., 2004)
Data3	Corrected SWE from airborne Gamma remote sensing	SnowModel assimilation	Derived from data at NSIDC (Cline and Carrol 2004)
Data3	Gravimetric soil moisture	Correction to Gamma SWE	Personally from K. Elder

Data Assessment

Data1a and Data1b: Many meteorological stations exist in the study area. In addition to established Federal monitoring networks such as the National Resource Conservation Service's Remote Automated Weather Station (RAWS) and Snowpack Telemetry (SNOTEL) networks, more ad hoc stations are also placed in the area. Transportation departments, avalanche forecast centers, ski areas, farmers, and individual research groups all have an interest in monitoring weather at particular locations. As Dr. Liston said, some of the data from these stations are archived; some are not; some are available; some are not. It is not always clear who owns certain stations. Given the primary objective of the study to create a legacy data set, Dr. Liston sought to get as much data in the area as practical—as much as "I can physically get my hands on." This is not a simple task. There is no comprehensive listing of meteorological stations, and Dr. Liston spent a lot of time searching on the internet and contacting colleagues. As a senior researcher familiar with the area, however, Dr. Liston already had a sense of what he was seeking. His primary criteria for selection were that the stations were located in one of the MSAs, that they provided relevant data with hourly or better temporal resolution, and, perhaps most importantly, that they were actually accessible (i.e. on the internet).

In addition to selecting individual stations, Dr. Liston chose to use only certain variables from each station. Precipitation was an especially important variable, but it is notoriously difficult to measure accurately, especially when frozen (see for example Doesken and Judson (1996)). Dr. Liston wanted the precipitation sources to be consistent across MSAs and to account for elevation differences. He, therefore, chose not to use precipitation data from any of the stations and use LAPS for the overall precipitation forcing. The LAPS precipitation data is not necessarily the most realistic, but it is consistent and considers differences in elevation across the study area. Because he includes a precipitation correction factor based on assimilation of the detailed SWE measurements from CLPX (Data3), he was not as concerned about getting realistic so much as consistent precipitation to force his model.

For the other four parameters, Dr. Liston chose to use all of them from some stations and only one or two (temperature and humidity) from others. This decision was partly based on the nature of the data provided from each station (resolution, available variables), but it was also

based on what sort of inputs MicroMet expects and Dr. Liston's personal knowledge of the individual stations and their location. For example, Dr. Liston did not use wind data from any of the SNOTEL stations. MicroMet assumes that the wind speed and direction inputs are fairly broad-scale forcings from top of ridge or top of canopy. MicroMet then reduces the wind in the canopy. Because Dr. Liston knew that SNOTEL stations are usually in small forest clearings (and often valleys) the wind direction and wind speed "don't mean very much".

The location of SNOTEL stations and their limitations tend to be fairly common knowledge within the snow monitoring community, but it may not be readily apparent to many atmospheric modelers. Dr. Liston has spent four-and-a-half years of his career in the field, and he emphasized the value of this extensive field research. "I'm a modeler, but I do a lot of field work. As a consequence, I avoid a lot of the pitfalls some modelers might fall into because they're not as familiar with the natural systems and the data sources they use."

Data2: For the vegetation and topography inputs, major assessment criteria were scale, coverage, and ready availability, but also the potential to reuse the data. These are common, well-known reference data sets that were of the highest resolution available for a study at this scale and coverage. The fact that they are reference data sets with national coverage makes them especially attractive because it is likely that Dr. Liston will be able to use them again in future studies and, therefore, will not need to redo any processing necessary to bring them into the model. This is an important point. Dr. Liston said, "I'm always looking for data sets that I can apply to other applications."

Data3: These data were collected as part of CLPX and were essentially designed to accommodate Dr. Liston's study (among others). There was no need to assess their applicability, per se, but Dr. Liston did consult extensively with co-authors Elder and Cline on how best to use the data.

Data Acquisition

All data used in this study were either available locally, sent to Dr. Liston by colleagues, or downloaded through FTP. Dr. Liston's general approach is to write simple scripts that go out and

download all the data and then process them into what he needs for his application. He doesn't like sophisticated web interfaces that allow subsetting, regridding, etc. because they don't provide a record of what he has done. "There is a trend toward these funky, GUI web interfaces, but they are not that useful to me." It's important to him to have a complete record of everything he did to process and display the data in the form of notes and non-proprietary scripts as well as the final results. This is important to the integrity of the study, but it also allows him to easily go back and repeat the acquisition process in the future if he needs to use the data again. When asked if there comes a point with some of his broader scale studies where the data volumes make this sort of bulk-get-then-process approach prohibitive, he said, so far, he hasn't reached that point. He noted that as we were talking, he was downloading 380 GB of data. It would take several days, but that was OK. Together with his colleague, Chris Heimstra, they have 25 TB of storage on their personal workstations, and "we're using every bit of it."

Data Preparation

Data preparation and preprocessing was a significant part of Dr. Liston's work.

Data1a: The data from the different meteorological stations contain different sets of variables and come in a variety of different formats with different time steps (e.g. 10 min., hourly, 3-hourly). Each of the six general data sources had their own standard way to format the data, i.e. some form of specialized ASCII. Dr. Liston needed to create scripts that extracted the relevant parameters, addressed missing values, aggregated the data as necessary into hourly values, and formatted them so they would run in MicroMet. He also applied quality control scripts that identified things like out-of-range or spurious values according to a standard methodology defined by Meek and Hatfield (1994). This customized scripting required significant effort, and no individual meteorological format was notably easier than any other to handle. "They all required their own specific attention." One particular issue was the use of tabs as a delimiter between values. This is common on PC platforms and with data exported from Microsoft Excel™, but because it is an invisible character it can be difficult to deal with especially in a cross-platform environment. Dr. Liston works on a UNIX platform, but cross-platform compatibility is important to him.

Data1b: Processing of the LAPS data is similar to the meteorological stations, but is generally easier to deal with because it is a consistent format that Dr. Liston has worked with before. The native format of LAPS is gridded netCDF, and while MicroMet can handle gridded meteorological inputs it cannot handle both point and gridded inputs at the same time. As a result, Dr. Liston needed to convert the 10 km grid to three 5 x 5 arrays of points covering each MSA. Each point is at the center of a LAPS grid cell and is handled by MicoMet in the same way it handles a meteorological station.

Data2: The National Elevation Dataset (NED) is available in a variety of raster formats, but historically the data has been geared toward GIS users. Co-author Heimstra has developed GIS routines to reproject and convert the data into the gridded format expected by MicroMet and SnowModel. Similar reprojection and conversion routines were applied to the National Land Cover Data Set. In addition, the land cover data are reclassified to the land cover classification used by MicroMet and SnowModel.

Data3: The two SWE data sets assimilated into the SnowModel were both collected as part of CLPX and are readily available from NSIDC (http://nsidc.org/data/clpx/), but Dr. Liston applied a correction to some of the Gamma data and the ground-based measurements were averaged over the 1 km^2 Intensive Study Areas. Liston et al. (2008) and Cline (2009) describe how having an accurate soil moisture measurement is essential to getting an accurate SWE estimate from the Gamma, especially for shallow snow. Dr. Liston applied a correction to the Gamma data for the North Park MSA during one of the observation periods based on the soil moisture measured on the ground as part of the experiment rather than the general background soil moisture used in the data set at NSIDC. Dr. Elder calculated average soil moisture for the North Park MSA from the hundreds of soil samples collected during CLPX, and Dr. Liston used this to apply the correction after extensive discussion with Dr. Cline, the Gamma expert. For the ground-based measurements, Dr. Elder calculated average SWE over each ISA for each IOP simply by multiplying the mean snow density calculated from the snow pit measurements by the average depths calculated from hundreds of depth measurements (Elder et al., 2009).

All of this data preparation was done with relatively basic tools, primarily UNIX shell scripts and Fortran77. Dr. Liston made a point of how he completely avoids proprietary software. This

is partly so he has an openly accessible record of everything he did to prepare and present the data. The scripts can then be used "by any graduate student in the world, and I have a lot of them." Furthermore, when using proprietary software "there will come a point where a program won't allow you to do what you want to do, and you can't fix it. It's a black box."

Summary

Dr. Liston brought together several models into a consistent modeling framework to simulate snow accumulation and distribution across three study areas with very diverse terrain and vegetation. A central goal was to produce the best possible high-resolution representation of distribution of snow water equivalent over the study areas for use by a variety of cold-land process studies. Early steps in the process were to identify, process, and QC meteorological inputs from many different sources. SnowModel then used these and other inputs such as vegetation and topography to produce initial SWE values, which were then improved through a data assimilation process. This was an involved complex study, but much of the effort was simply dealing with the very disparate meteorological data from different weather station networks. The complete lack of standardization in the meteorological data presented a significant hurdle for Dr. Liston in this and other studies. He used primarily Unix shell scripts and Fortran programs to prepare the data, and the free Grid Analysis and Display System (GrADS) to visualize the model results.

The Multiple Element Model

Overview

The Multiple Element Model (MEL) is an expansion of an earlier model developed by Rastetter and Shaver (1992). It is used to compare interactions between the cycles of any two ecosystem elemental resources. It has typically been used to compare the interaction of nitrogen (N) and carbon (C) cycles to, for example, better understand how N limitation will constrain vegetative responses to increased CO2 (Rastetter et al., 1997; Rastetter et al., 2005). The model is defined at a plot scale (e.g., a described forest or experimental plot), so it can be parameterized by known state variables and fluxes of C and N. Rastetter and others have typically

parameterized the model using data from deciduous northeastern forests (Rastetter and Shaver, 1992; Rastetter et al., 1997; Rastetter et al., 2001; Rastetter et al., 2005), but the model has been extended to other ecosystems, including Arctic ecosystems (Herbert et al., 2004). Dr. Rastetter notes that a key issue is capturing sufficient parameterization data (i.e. N and C stocks and fluxes) from different plots or ecosystems.

Dr. Rastetter's use and development of the model has evolved over time. While this case study specifically examines the work conducted for Rastetter et al. (2005), it necessarily explores the work Dr. Rastetter has done in a series of papers since his first publication of the model in 1992, especially Rastetter et al. (2001). Therefore, the overall workflow description in this case study applies somewhat generically to Dr. Rastetter's overall work while using the data used for Rastetter et al. (2005) as illustrations. Dr. Rastetter made the point that he is a theoretical modeler, and that most of his papers address the idea that "there are a number of things about ecosystems that we don't have a good handle on." He, therefore, runs a series of simulations to answer the question "How important is it that we get a handle on a particular process?" In the case of Rastetter et al. (2005), he sought to answer the basic question "Does it make a difference [in a global warming scenario] if the nitrogen loss from the ecosystem is in a recalcitrant, organic form unavailable to plants and microbes versus a labile, inorganic form that is available to plants and microbes."

Dr. Rastetter ran a series of simulations with different ways of modeling nitrogen loss from the ecosystem. Building from Rastetter et al. (2001), he parameterized the model primarily using data from the US Forest Service Hubbard Brook Experimental Forest—a long- and well-studied, even-aged, second-growth forest in central New Hampshire composed of about 80-90% hardwoods and 10-20% conifers (Hubbard Brook Ecosystem Study, 2001). All the simulations suggest that it makes a big difference in ecosystem response to warming and increased CO_2 whether nitrogen loss from the system is Dissolved Inorganic Nitrogen (DIN) or Dissolved

Organic Nitrogen (DON),[1] but the differences are not evident for sixty to one hundred years. DON losses can have significant impact on long-term C sequestration in forested ecosystems. Of relevance to the Arctic, Rastetter et al. (2005) note that the effects of DON loss would be masked in an environment such as the tundra which is becoming more woody in response to elevated CO_2 and warming (Sturm et al., 2001).

Figure 4.5 outlines the general process Dr. Rastetter went through in this and related studies and shows where data were used during different steps. Table 4.4 provides more specifics on the data. The steps in Figure 4.5 are as follows.

1. As with any study, the first step is to set out the science question. This step is important to call out in this study, because it can be related to what data are available.
2. Decide how to represent the necessary processes in the equations to address the question. In this case, Dr. Rastetter represented the N loss in four different ways related to the C:N ratio.
3. Adjust standard model. MEL varies from the "standard" N uptake model (Vitousek et al. 1998) to adequately capture necessary processes to describe DON loss and increased N demand by the ecosystem in response to elevated CO_2 and to reflect that N uptake by plants and microorganisms and N loss all occur simultaneously. The standard model assumes a sequential progression where microorganisms immobilize all the N they can, then plants take up what they can, then loss may occur. This simplifying assumption arose partially because it allowed the model to be formulated only on net and not gross

1 "Dissolved Organic Nitrogen (DON) is defined as the difference between Total Dissolved Nitrogen (TDN) and Dissolved Inorganic Nitrogen (DIN), which comprises nitrate (NO_3-), ammonia (NH_4+) and nitrite (NO_2-). DON is not a single compound but a mixture of compounds ranging from simple amino acids to complex humic substances. Most studies on nitrogen loss from ecosystems measure only the inorganic forms of nitrogen, ignoring mobile organic forms such as amino acids, aminated sugars, and humic acids that dissolve into soil water and can be lost as such" (Barbero, 2006). In Rastetter et al. (2005), they simplify their analysis and define DON to only represent forms of N unavailable to plants and microbes, recognizing that there is growing evidence that plans can access some DON.

mineralization, which is much harder to estimate.[2] Dr. Rastetter needed to remove this simplification because a central purpose of the study was to better understand how plant uptake, microbial immobilization, and N loss differ if N losses are DIN or DON.

4. Run the simulations with high and low DON loss ratios and with the different N loss formulations. All simulations were forced with doubled CO_2 and a 4°C temperature increase, as predicted for New England by the IPCC (2001), and allowed to run to a new steady state.

5. Review results. The investigators used the basic Rastetter and Shaver (1992) version of the model to predict plant and soil C and N stocks, calculate differences from original values, and to partition the changes in total ecosystem C according to different factors.

6. Modify parameterizations as necessary. While it was not the case in this situation, Dr. Rastetter noted that sometimes the process of running the model uncovers inconsistencies in the data. He gave one example where he used the Hubbard Brook Experimental Forest data set to see if he could reproduce the growth of the forest since the last glaciers left 14,000 years ago, but he couldn't get the system to grow fast enough in the model. It turns out that weathering rates estimated for phosphorus (P) were off. When he presented his results at a meeting at Hubbard Brook, he discovered that people were in fact doing research on how plants may be fostering more rapid break down of minerals. This could lead to a correction in the P weathering rates. In another example, he and a colleague realized that they had misinterpreted a data value. The point is that there can often be a recalibration process when running the model.

[2] Mineralization converts organic N to inorganic N. Net mineralization is the total or gross mineralization minus what has been immobilized (converted to organic form) by microbes. Net mineralization considers ammonium (NH_4+) and nitrate (NO_3-). Gross mineralization refers only to ammonium (NH_4+).

7. Present results. Dr. Rastetter emphasized how a major part of his job is communication. As a theoretician, he finds it especially important to communicate with empiricists. He encourages people to design experiments to empirically test his results and further understand processes. He even scoped out such an experiment in the conclusion of Rastetter et al. (2005). Of course, the results also often lead to new science questions to be assessed.

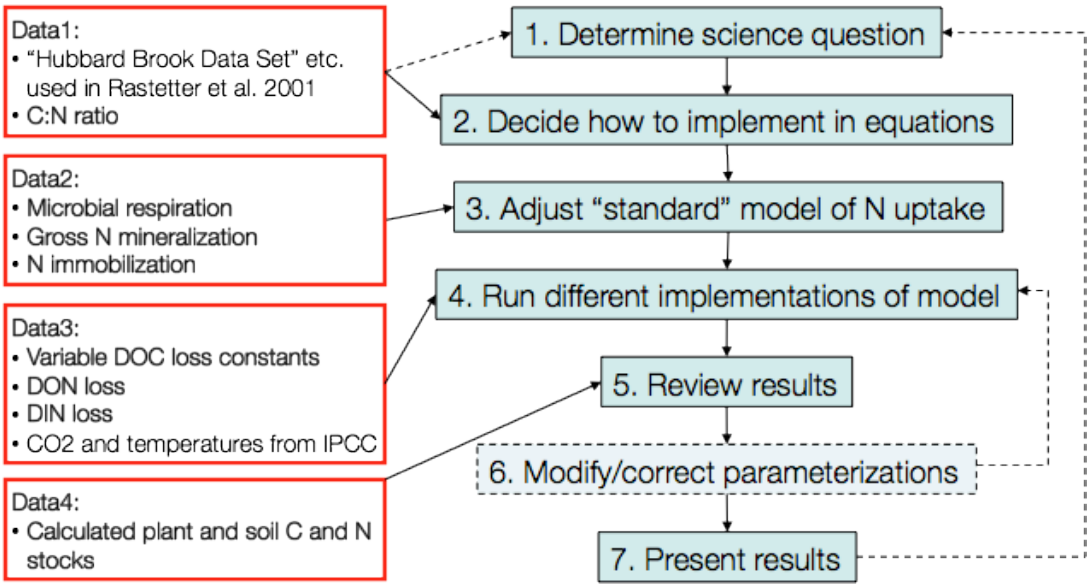

Figure 4.5. General workflow and data inputs for the research process leading to Rastetter et al. (2005) and similar studies.

The data used in this study are different than those used by Drs. Holland and Liston. The model does not have an explicit spatial dimension. The primary issue is developing the right parameterizations for the type of ecosystem being represented. Data are not continuous temporal or spatial fields but are typically single-value estimates of chemical stocks or fluxes. These values may be estimated from detailed allometric measurements or proxy measurements like using stream chemistry to estimate N loss. Values may also be calculated by different models and experiments.

Table 4.4 lists the different data used in the study. Data1 are the primary stocks and fluxes and basic parameterizations. The Hubbard Brook Data Set (In particular: Bormann and Likens, 1979; Whittaker et al., 1979) with some updates, notably to fine root dynamics by Fahey and

Hughes (1994), provide most of the stocks and fluxes. Other sources, including personal communication, are fully cited in Rastetter et al. (2001) Table 1, and Table 2 shows the parameters estimated by model equations or other cited approaches including an earlier MEL study (Rastetter et al., 1997). Goodale et al. (2000) provide the basic C:N ratio necessary to develop different model scenarios for DIN and DON loss. The Hubbard Brook data are central to the whole project. They are necessary to run the model in a way that approximates a specific ecosystem and enables the inclusion of more detailed processes. As such, the data enter the process very early on to help define how specific processes can be defined mathematically. They may even inform the sort of scientific questions that can reasonably be assessed with the model (This is indicated with the dashed line in Figure 4.5 and is discussed in more detail below). Data2 are changes in model values that are calculated to maintain a steady state when the model is expanded to consider gross and net mineralization processes. These are not observed values but calculated calibrations based on well-described theory. Data3 are different constants derived from the literature to parameterize the alternate ways of modeling DOC loss and how this is associated with DON loss. Data3 also includes the global warming forcing values from the Intergovernmental Panel on Climate Change (IPCC 2001). Data4 are also calculated values that allow the investigators to partition the changes in ecosystem C into soil and plant components for more detailed analysis.

Table 4.4 Data used in Rastetter et al. (2005) as shown in Figure 4.5. See extended table in Appendix C.

Stage	Data Collection/Data Set	Application	Data Source
Data1	Ecosystem stocks and fluxes and estimated parameters primarily from the Hubbard Brook Experimental Forest.	parameterization, forcing	As described in detail in Rastetter et al. (2001) Tables 1 and 2..
Data1	C:N ratio	parameterization	Goodale et al. (2000)
Data2	Microbial respiration	parameterization	Calculated modification to values in Rastetter (2001) to maintain assumption of steady state.
Data2	Gross N mineralization	parameterization	Calculated modification to values in Rastetter (2001) to maintain assumption of steady state.
Data2	N immobilization	parameterization	Calculated modification to values in Rastetter (2001) to maintain assumption of steady state.
Data3	Model 1: constant DOC loss	parameterization	Baseline assumption.
Data3	Model 2: constant: proportional to organic matter in the soil	alternate parameterization	Based on values from Neff et al. (2000).
Data3	Model 3: constant: proportional to C:N ratio	alternate parameterization	Based on values from Aitkenhead and McDowell (2000).
Data3	Model 4: constant: proportional to microbial respiration	alternate parameterization	Based on values from Brooks et al. (1999).
Data3	2x CO2 and 4°C temperature increase	forcing	IPCC (2001) (for New England)
Data4	Calculated plant and soil C and N stocks	analysis	Derived from original MEL (Rastetter and Shaver, 1992).

Data Assessment

To understand Dr. Rastetter's approach to data assessment and use, we must consider his research context. As mentioned, Dr. Rastetter emphasized that he is a theoretical modeler. His goal is to better understand ecosystem processes and their relative importance, not to make specific predictions of a future state. As he put it, his approach is "modeling for understanding" not "modeling for numbers." In this context, precise data are somewhat less important. In a situation where you are trying to make a precise prediction, like CO_2 in the atmosphere in 2100, then numbers make "a heck of a lot of difference." If you are looking at a more qualitative question, like whether the form of N loss affects ecosystem response to CO_2, then the numbers "don't make that much difference." Dr. Rastetter said that he could have taken a more abstract approach to this question, like the approach initially used in Rastetter and Shaver (1992), but that would limit his audience. "Then I'm just talking to theorists, and that just drives me nuts." Dr. Rastetter feels that it is critical to collaborate and communicate with empiricist scientists. To help that communication and to facilitate publication in journals beyond strictly modeling journals, he needs quality data. Data become central to the communication between theoreticians and empiricists. People can better relate to the argument and internalize the concepts if they can tie things down to a real ecosystem. They "need numbers they can feel comfortable with" and the reviewers "want justification for every number that you use." So while precise data may be less important to a theoretical argument, authoritative data are essential for Dr. Rastetter to publish widely.

Data1: The Hubbard Brook Experimental Forest was established in 1955. The associated Hubbard Brook Ecosystem Study began in 1960 as one of the first comprehensive studies of an entire ecosystem (Hubbard Brook Ecosystem Study, 2001). The study has developed one of the most comprehensive data sets on ecosystem fluxes and element stocks. Dr. Rastetter called it "one of the great data sets." In terms of assessing what data to use in this application, there are few options. Dr. Rastetter mentioned the work of Phillip Sollins at the H. J. Andrews Experimental Forest in the western Cascade Range of Oregon and the work of Gaius Shaver at the Arctic Long Term Ecological Research Site on the North Slope of the Brooks Range in Alaska, but the Hubbard Brook Data Set has the distinct advantage that Fahey and Hughes (1994) have done excellent work to characterize the dynamics of fine roots at Hubbard Brook.

This sort of work has not been done elsewhere, and Dr. Rastetter said a major data need for him is "everything below ground". In addition, Goodale et al (2000) also worked in Hubbard Brook and the overall White Mountains to provide the C:N ratio information necessary to asses Dr. Rastetter's N loss question. Dr. Rastetter said he is always looking for ways to get the model in a form that he can publish (i.e., supported by authoritative data), so he is continually searching for suitable data across different sources and different ecosystems and trying to stitch things together. In this way, data availability can affect how Dr. Rastetter can formulate specific questions in his model and can even have some effect on the form of the particular questions Dr. Rastetter tries to address (dashed arrow in Figure 4.5.). As a theorist, Dr. Rastetter does not let data availability determine the science questions he seeks to answer, but he admits that publication pressure and the need to effectively communicate with his "empiricist friends" can push him in certain directions.

Dr. Rastetter is very conscious of this interplay between data and models, especially in parameterization and calibration. While precise values for a particular flux, say, may never be known, one must recognize that when calibrating a model to fit a specific number, one subsumes not only the uncertainty in the model structure but also the uncertainty in the measurement. Sometimes, that uncertainty can even be a bias. Dr. Rastetter gave the example of how the work by Fahey and Hughes (1994) changed the earlier N uptake rate by more than 50%. Further, on a more theoretical level, Dr. Rastetter notes that model construction is often driven by how data are actually collected: "the model is a representation of how we measure the system more than it is a representation of the system. There are two levels of separation." These are fundamental issues in modeling, and they play a strong role in Dr. Rastetter's work. It is interesting to note how on one hand, as a theoretician, Dr. Rastetter has little need for precise data. On the other hand, he is acutely aware of the need for authoritative data across the ecosystem for his results to be accepted, understood, and applied. In this sense, he prefers data that have been published in the peer-reviewed literature. It is not his job to review the accuracy of different data, so it is good that "someone has checked it over." All these considerations led Dr. Rastetter to the Hubbard Brook Data Set for this and previous studies.

Data2. These data are calculated values not observations. Their "assessment" is inherent in how Dr. Rastetter chose to describe the N processes in the model equations. They are used in a form of parameterization and calibration, though, so they are worth calling out given the uncertainties of parameterization and calibration.

Data3. These data are what was necessary for Dr. Rastetter to run the multiple implementations of his model. He sought to avoid the potential criticism that he didn't consider how N is lost from the system. He tried to consider approaches representative of the current work in the community. So assessment was a matter of identifying accepted sources in the literature. The other parts of Data3 are the CO_2 and temperature forcings using benchmark IPCC values.

Data4. These data are also calculated values. Similar to Data2, they illustrate how model results can themselves be a form of data.

Data Acquisition and Preparation.

Most of the data used by Dr. Rastetter are published numbers, so acquisition and preparation is simply a matter of transcription and interpretation. All the Long Term Ecological Research sites like Hubbard Brook have data access web sites. Dr. Rastetter says he uses these systems occasionally, but that her prefers something from the peer-reviewed literature. Data preparation is usually just a matter of unit conversion. Sometimes, preparation is somewhat of a research question, like figuring out how to divide a single number into different stocks with different fluxes. Sometimes data may need to be corrected for limitations in measurement techniques. For example, most people report the total extractable NH_4, which is NH_4 that is extracted from soil using a strong potassium chloride solution. This does not distinguish between the NH_4 in soil solution and the NH_4 stuck to soil particles. This ratio can have significant impact on the long-term uptake of NH_4 by the ecosystem.

Summary

Dr. Rastetter is a theoretical modeler. In this application he sought to improve understanding of how the form of nitrogen loss from an ecosystem impacts ecosystem response to warming.

Data used in this study are quite different from the earlier case studies. The primary need for observational data is to parameterize the model. These data are typically single values of element stocks or fluxes published in the literature. There are few high-quality data sets that describe all these sort of parameters for a particular ecosystem, but having good parameterization is essential to Dr. Rastetter's work and collaboration with empirical researchers.

Dr. Rastetter's work provides an interesting contrast to that by Dr. Holland and Dr. Liston. Dr. Rastetter not only works in a very different domain—biogeochemistry vs. cryospheric science—but also with a different perspective—that of a purely theoretical modeler. The next chapter explores these contrasts as well the commonalities across the case studies in a cross-case analysis.

Chapter V
Analysis and Results

This chapter presents analysis of each individual case followed by a cross-case comparison and synthesis. Each case is examined independently for possible areas of data center and data system improvement, but the analysis of later studies inevitably builds on the earlier work to create a cohesive narrative. The areas for data center improvement that emerge are italicized in the text. The comparison and synthesis then seek to create possible principles and practices for data managers to improve Arctic system modeling processes. The central proposition is that there are instructive themes in how different modelers assess, acquire, and prepare data for their models. A goal is to suggest data management techniques or requirements for data systems to improve access and use by modelers.

Dr. Holland and the Community Sea Ice Model

One of the first things that became evident in this initial study is that modelers have a different conception of data than do most scientific data centers. It was telling that the first "data" input into Dr. Holland's process were model outputs from the CCSM and other GCMs. This is not the sort of data held by most data centers, which tend to focus on what might better be called "observations" or "measurements." There are data centers, such as the CMIP3 archive at PCMDI, who handle model output, and it is increasingly an issue many data centers will have to face, but for the most part, climate related data centers focus on satellite, aerial, and shipborne observations and field measurements. (The rest of this thesis will use the general terms model outputs and observations to distinguish between these broad data types.) Greater consonance in the management of these different data types could help climate modelers significantly. Dr. Holland noted how the detailed standardization of model output through the CMIP3 project really enabled more extensive science. If observations were more coherent with CMIP data formats, grids, and naming conventions, there would be even greater benefit. In short, *CMIP has created a set of standards for data centers to consider when presenting certain data—notably broad-scale, gridded, time series of climatic variables.*

In a sense, it comes down to convenience. Dr. Holland readily used data she knew to be less accurate for her initial comparisons simply because it was easier. She only used the "better" data

set when she needed a very thorough and robust analysis. Convenience can be seen as a way to reduce effort. This balance between convenience and data quality or uncertainty is something modelers constantly need to consider, so data centers can clearly help by making their data more convenient to access and use.

We can consider convenience as reduced effort, and we can qualitatively measure the effort of a research study as a series of steps or decisions points. So by reducing steps in the process, we have at some level made it more efficient, more convenient (cf. Pressman and Wildavsky, 1973). To illustrate this, consider if NSIDC had provided its sea ice data in netCDF-CF (ideally in a latitude-longitude grid in accordance with CMIP standards). Dr. Holland could have essentially removed the initial assessment process (Figure 4.2, step 3) and several decision points. In this study that might have saved a day or two of effort, estimates Dr. Holland. In other studies, with greater use of observations, it might be several days of effort saved. Of course, there is not one format to serve all communities, but for the GCM community, there is a working standard that continues to develop through the CMIP process. *Ultimately, data centers should provide data in multiple standard or common formats.* NSIDC has, in fact, recently introduced a new data access system that enables search, subsetting, and delivery of data in multiple formats and grids for major gridded data sets.

In addition, to making the data more convenient in format and grid, Dr. Holland desired tools to enable cursory data analysis of data before actual data acquisition. The Sea Ice Index provides some of this in the form of browse images, time series plots, and comparisons to climatologies, but they are relatively static and do not allow the user to change any of the parameters or variables. It is important to note, however, that while Dr. Holland expressed interest in these types of tools, she typically gets her data through basic ftp. Further, she does not look to data centers to provide information on data quality or applicability. She prefers instead to talk to colleagues and get specific recommendations. Ideally, she would like one clear and definitive product. This has long been an issue for sea ice, in particular, where diverse sea ice data products have different strengths and weaknesses depending on the particular application (Meier et al., 2001; Parsons and Duerr, 2005). Nevertheless, where possible, *data centers, in collaboration*

with scientists, should clearly and succinctly indicate which products are most suitable for different applications.

This particular study by Dr. Holland (Holland et al. 2006) was a very specific case on the use of the CSIM and the CCSM more generally, yet it revealed some interesting insights. An examination of other applications of the CCSM could provide additional lessons on this important class of models. To use the language of the Open Archival Information System Reference Model (ISO, 2003), *CSIM and CCSM modelers are a specific "designated community" to be explicitly considered for key products held by geophysical data centers like NSIDC.*

Dr. Liston and SnowModel

Unlike Dr. Holland, Dr. Liston's objective was not to predict possible future states, but rather to characterize the current state of a very complex parameter–snow–over variable and complex terrain. His study, therefore, included much more direct manipulation and use of observations. He even commented that he made extra effort to get more and better data for this study than he might do for others because the goal of this study was to produce a benchmark data set. This relates to the issue of convenience discussed above. Dr. Liston says that gathering forcing data is a constant assessment of payoff vs. effort. In this study, he was willing to work hard to get high-quality data. That was a central purpose of the whole Cold Land Processes experiment. In other situations, where he was not producing a data set but perhaps examining a particular process, the balance between payoff and effort would be different. The effort, in this case, was largely finding and then preparing the meteorological forcing data from the various meteorological stations.

As with Dr. Holland, much of the data preparation effort was driven by the differing data formats and conventions, but in this case, there is no obvious standard on which to converge. Data come from diverse local, regional, and federal sources, each with their own established conventions. SNOTEL data, for example, are still in English units. Changing or harmonizing these conventions requires significant change to established social and technical infrastructures, which in turn creates large implications and tensions (Edwards et al., 2007). Further, the use, preferred formats, time-steps, terminology, and units of meteorological data are so diverse as to

make the creation of community built standards daunting. *Data managers can endeavor to be honest brokers, and provide data in multiple formats.* Indeed, with this sort of point data and diverse user base, it may be most useful to provide *a mechanism for users to create their own format on demand through a basic queryable, web-services interface to a database.* Continued research in and application of advanced semantic techniques could also be fruitful.

Dr. Liston provided one anecdote that starkly illustrates how much time could be saved with more consistent data. He described a master's student who was working on a similar application of SnowModel and had collected data from about 12 meteorological towers from multiple networks in Oregon (NRCS, NWS, LTER). Unlike Dr. Liston, the student was not familiar with the data type and was not well versed in Unix scripting. It took him about three months to figure everything out and get all the data together for the model. Dr. Liston estimates that with his experience, it might have taken him about three weeks, but if the data were all in a consistent format, it would have taken both the student and his mentor only three days. This suggests format harmonization provides significant time savings across science given the myriad applications of these data.

The term data format should be clarified in this context. Most meteorological station data are in an ASCII format, but that is a very general characterization. A more precise specification is what Raymond (2004) calls the metaformat and is what most people consider to be formats. Examples for ASCII include delimiter-separated values and XML. At a greater level of specificity is what the NASA Strategic Evolution of Earth Science Enterprise Data Systems (SEEDS) Formulation Team (2003) calls the format profile. This is a specific implementation of a metaformat and could include more syntactical details such as the order of columns in a tabular data set or machine-specific considerations such as byte order and 32 vs. 64-bit words for a binary array. It is at this level of the data format profile where most problems occur. The devil is in the details. These details can be a syntactical issue like which delimiter to use between columns (a real issue in CLPX (Parsons et al., 2004)) to semantic details like what exactly is meant by "temperature." These issues are not unique to meteorological station data, but they apply to many regular, point-based, field measurements of the environment. Harmonizing data profiles, conventions, vocabularies, etc. within disciplines, let alone across disciplines, is a

significant undertaking for both the data management and scientific community. This is a topic for more research, but in the short term *data managers can highlight the issues and facilitate community efforts to harmonize formats and conventions.* Fetterer (2009), for example, has been coordinating an effort to harmonize sea ice observations.

While being flexible in how they deliver data, data centers must also be consistent in how they serve their data over time. Dr. Liston made the point that he is always looking for data that he can reuse. "If I can acquire a data set that fits all my projects, that's what I want to do." This implies that data centers should be conservative in making changes to data formats, grids, etc. It is OK to add more formats, but old ones should not be removed lightly. When data do change, such as with new calibrations, this should be clearly indicated. These data provenance issues can be all the more challenging when providing data through a queryable database. While a resource, such as a central database, that delivered all the meteorological data in one format would have saved Dr. Liston a lot of time, he also made the point that he was not very interested in sophisticated web interfaces because he wanted a precise record of what he did. *Data centers should track and include with the data a record of any sort of processing, such as subsetting or regridding, that is done prior to data delivery.*

Overall, Dr. Liston did a lot of basic preparatory work that could be done by data centers, be it standard error correction algorithms on the meteorological station data, interpolating and regridding the land cover data, or making point values out of the gridded LAPS data. The question is whether they should. Providing these kinds of services, could be beneficial to some users, but scientists can be rather traditionalist. Dr. Liston uses a lot of data, but he gathers it all together on his workstation where he does all the processing in a way that he can readily track and record with basic scripts and annotations. This gives him precise control and a complete record of everything he has done, and he has been working over time to make his tools and processes reusable. Data centers today have strong pressure to innovate and provide useful tools to manipulate large volumes of data, but at the same time there is a pressure to be more efficient, and to develop sustainable, broadly supported data systems. While making data more consistent increases scientific efficiency, it is not clear whether providing additional pre-processing services always will increase efficiency. More research is needed on the costs and benefits of centralizing

certain data services. The answers are likely to be very different for relatively small in-situ collections vs. large and growing model output and remote sensing observations.

Another area where a data center role is unclear is in documenting uncertainty. While it is clear that comprehensive documentation of uncertainty is desirable, it is not clear whether it is always possible. For example, Dr. Liston made careful and educated choices in his selection of variables from the various meteorological stations. He made his decisions based on his personal expertise and field experience, but also on common knowledge in the snow community about the site characteristics of the stations. *To enable broader, more interdisciplinary, and systemic data use, data centers need to try and capture and convey this sort of tacit community knowledge*. It is natural that much scientific knowledge is exchanged through interpersonal communication. The goal, therefore, should be to make that knowledge exchange more broadly accessible. Social networking tools seem like an obvious approach. It has often been suggested at scientific data management workshops and elsewhere to use the comment and discussion tools used by retail web sites or similar to discuss and annotate data. Few, if any, of these schemes have been implemented, though, and it is unclear if there is the critical mass of participants necessary for such social networking approaches to succeed. More experimentation with these tools is warranted. In the longer term, semantic research and ontology development promise to help convey uncertainty in rich and formal way.

Dr. Rastetter and the Multiple Element Model

This study was very different from the previous two. It was in a completely different scientific disciple, using a very different approach. The model used very different data and no data from NSIDC or other data that I was familiar with when I began the study. Moreover, as a theoretical modeler, Dr. Rastetter has a very different view on data. As a result, there are fewer obvious lessons to be learned on specific data management practices and more to be learned at a theoretical level. A central theme is that data are central to communication between theoreticians and empiricists. The data that Dr. Rastetter use are almost all for parameterization, and he is very aware of the sensitivities involved. Parameterization can been seen as a process of intelligently determining that certain parts of the system can be neglected in the model and certain parts can be represented through semi-empirical or imprecise mathematical formulae. Parameterization is

a necessary but fraught part of modeling (McGuffie and Henderson-Sellers, 2005). Further, "The most advanced parameterizations have theoretical justification." (McGuffie and Henderson-Sellers, 2005). Parameterization, therefore, becomes a focal point for that scientific conversation between empiricists and theoreticians.

Dr. Rastetter emphasized how it was important to have authoritative data, particularly something that has been peer-reviewed and that he could cite. Because he is typically using single number values for element stocks and fluxes, the numbers can be found in traditional peer-reviewed literature. For other types of parameterization, and especially when getting into data assimilation like Dr. Liston, the data are more complex and less likely to be formally published in conventional scientific literature. That does not necessarily mean that these data, typically held by scientific data centers, cannot be properly cited and even peer-reviewed.

Data citation has been described in the literature (e.g., Costello, 2009; Klump et al., 2006), and many geophysical data centers, including most NASA centers, recommend specific ways to cite their data but their approaches vary. Some data centers, including NOAA National Data Centers, do not request formal citation and simply request data be acknowledged in the text. Some data centers, including some USGS centers, take different approaches for different products. For example, citation may be requested for digital maps while only acknowledgement may be requested for tabular data. Occasionally, a data publisher may request that data users cite a journal article or other document describing the data. Ironically, these types of citations seem to be most broadly used despite the fact that the citation does not directly refer to the actual data used. In some cases, the data may actually be a supplement to the article, but more often the data extend well beyond a specific article.

In the Arctic, the International Polar Year explicitly recommends data citation in its Data Policy (http://classic.ipy.org/Subcommittees/final_ipy_data_policy.pdf) and has developed guidelines for how data should be cited (http://ipydis.org/data/citations.html). These guidelines, like any, are imperfect, but they harmonize different approaches and have been adopted by data centers around the world. *Data centers can and should use these guidelines now to indicate how their data should be cited in a way that gives fair acknowledgement of the data author.* These guidelines can then serve as a basis for evolving approaches to formally cite data.

Data citation, while helpful, is only the beginning in asserting the authority of a data set. Citation needs to be coupled to some sort of quality assurance or peer-review scheme. In many ways, good data have always undergone some level of peer-review, and many NASA and NOAA data centers vet the data they handle, but there is no formally recognized or established process. Developing that process is a greater challenge than data citation, but it is no less vital to modern, data-driven science. It is likely that data publishers will play an important role in establishing appropriate peer review processes. Community best practices could be established addressing some of the issues discussed in this thesis, such as standard formats and data validation, as well as more complex community issues such as determining what level of assurance is necessary to apply at large scales when millions of data files may be produced? For example, is academic review of processing algorithms such as documented in NASA's Algorithm Theoretical Basis Documents sufficient? It is rigorous, but does it receive the same recognition as peer-review? How does this contrast with review processes for research data collections produced by individual investigators or small projects that rarely produce the level of documentation or undergo the levels of review of the large programs? These, along with data citation, are the sorts of issues the data management community needs to address in collaboration with scientific researchers (Parsons et al., 2010).

Cross-Case Comparison and Synthesis

When comparing these different studies, what is immediately striking is the differences between the studies. These differences reemphasize how important it is for data managers to understand their audience and their needs. If data managers can better define their designated community and their data application, they can better target their user-needs-assessment efforts. These case studies suggest that there are certain categories of modelers and data application that may be instructive when assessing modeler data needs. These three modelers had six general data applications (Tables 4.2-4.4): forcing, assimilation, parameterization, assessment, analysis, and correction or calibration. These different applications create different data assessment criteria. For model forcing and data assimilation, consistency of the data is very important. For example, Dr. Liston’s variable selection of wind from the different meteorological towers was largely geared toward ensuring the data were consistent. For parameterization, the data need to be broadly accepted and authoritative. For example, Dr. Rastetter’s efforts to engage empiricists

hinged on using accepted data for the parameterization. For assessment of model results, the data need to be broadly representative or "true". For example, Dr. Holland made extra effort to get the more representative sea ice data for her full assessment.

Other categories of use would likely emerge in additional studies These could include validation or verification and model conceptualization. Dr. Holland was essentially doing verification when assessing how the CCSM output compared to the observations. Dr. Rastetter showed how data availability can influence how a model is conceptualized or mathematically defined.

One can also consider the types of modeling being conducted. Dr. Holland's work might be considered predictive modeling. As such she is largely interested in getting the best data representation she can of the predicted value, sea ice, to assess and verify her results. Dr. Liston might be considered a descriptive modeler, who is therefore interested in getting the most comprehensive and consistent inputs he can in order to produce the best overall description of the current state of a variable, snow. Dr. Rastetter as a theoretical modeler interested in understanding different ecosystem processes needs authoritative data in order for his theoretical arguments to be accepted by the broader scientific community.

Other ways of classifying models and their applications could also be identified. For example, Serreze and Barry (2005) define seven general model types used in the Arctic. The point is that by better defining their audience, data managers can better serve their needs. For example, NSIDC could benefit from a closer examination of predictive GCM modeler needs.

Despite the differences in these case studies, there were some commonalities. Chief among these is how modelers have a much broader conception of data than many scientific data centers, which tend to focus on observations from relatively narrow disciplines. Data centers cannot provide all the data needs of modelers. The literature and the models themselves are but two other data sources. Data centers, therefore, need to focus on what they are best able to provide to meet modeler needs. For example, in NSIDC's case providing consistent climate data records of key cryospheric variables in accordance with CMIP standards could greatly assist the GCM

community. Greater harmonization of disparate point measurements could significantly help with descriptive modeling and data assimilation.

Other commonalties also emerged from these studies. Table 5.1 lists some of the potential commonalities identified during the conduct and assessment of the three case studies. They are listed roughly in the order of the strength of the commonality as measured by how true each statement is for each model.

Table 5.1. Common issues or attributes across the three case studies. The upper case T in the columns on the right indicates the statement is very true for that model or case; the lower case t indicates the statement is somewhat true; and blank indicates that the statement is not true or it could not be determined.

Commonality	CSIM	Snow Model	MEL
1. Data availability limits types of studies.	T	T	T
2. Continually evolved and reapplied model over time.	t	T	T
3. Need to reformat data to input into model.	T	T	t
4. Data from data centers is acquired in bulk via ftp rather than through specialized interfaces.	T	T	t
5. Much of the data used does not come from data centers.	t	t	T
6. Multiple runs of different models are conducted (to satisfy reviewers).	T		T
7. Need to regrid data to match model.	T	T	
8. Need to be able to cite data sources (esp. for parameterization).	t		T
9. Spend a lot of time searching for data.		T	t

In reviewing these commonalities, we can begin to see some larger themes. It is not surprising that data availability can limit what scientific work is done (#1). Much of the necessary data are simply not available, but this limitation could also indicate a data discovery issue given that data centers are not necessarily a primary source of data (#5) and that two modelers called out discovery issues even though data discovery was not part of the study (#9). It is also not surprising that modelers build on previous work and continually evolve their modeling system (#2). This evolution suggests, however, a need for data to be consistently available over time in the same location, format, grid, etc. The need for consistency pairs with the common need to reformat (#3) and sometimes regrid (#7) data to emphasize the theme of convenience discussed earlier. Considering the issues of discovery and convenience suggests a need for data centers to identify and prepare specific products for specific community needs.

Careful examination of these themes can reveal more specific or actionable propositions that can be fully tested against the evidence in each case study. Table 5.2 shows how major propositions match up to the case study evidence. Sources of evidence include notes, recordings and follow-up on the interviews (I), the papers of each modeler (P), other literature (L), and the data and interfaces used by the modelers (D). Appendix C provides additional information on the data sets. An additional source of data is ethnographic or participant observation. In other words, the propositions are also tested against personal, professional experience and careful observation as an active participant in the Arctic science community. This is discussed in more detail in Chapter 6.

Table 5.2 lists some major propositions and shows where specific evidence for each assertion can be identified. All of these propositions are worthy of greater consideration by the data management community regardless of how well they are supported by these particular studies. Chapter 6 explores these propositions further and begins to develop a few more overarching principles.

Table 5.2. Some of the more viable propositions for data systems improvement that were tested during this study with indications of where there is evidence to support each proposition.

Proposition	Holland				Liston				Rastetter				Ethnographic Evidence
	I	P	L	D	I	P	L	D	I	P	L	D	
Data should be available in multiple common formats or specified on demand	x	x		x	x	x	x	x			x	x	Strongly supported.
Do not move/change data	x				x								Strongly supported.
Include provenance with data					x								Seems an obvious good idea, but there is little apparent demand for this or discussion of the topic. Other issues appear more pressing.
Provide recommended data citation					x			x	x	x	x		Strong evidence in data management community. Less in the scientific research community.
Data need peer review	x				x			x	x	x	x	x	A nascent but growing discussion topic.
A common criterion is the effort needed to use data--modelers want to simply download data in the same format, same grid, scalable, etc. (This requires understanding of specific needs.)	x				x	x		x	x				Strongly supported.
Focus on data not system--Content first! • Harmonizing data formats would reduce processing steps and save time • Providing multiple formats would reduce processing steps and save time	x	x		x	x		x	x	x	x	x	?	Strongly supported.
Data centers should clearly indicate the authoritative products for certain applications (parameterization, GCM validation, etc.)	x		x	x	x		x	x			x	x	Periodic requests for this, but it's difficult to match many products to diverse communities.

I=interview P=papers L=other literature D=data themselves

Chapter VI
Summary and Discussion

The Arctic environment is rapidly changing. To understand this change and its implications requires an integrated approach that considers the Arctic as a complex system. Models of various scales and types are major tools for this integrated approach. For the models to be effective and meaningful, they need quality data for forcing, parameterization, calibration, and assessment. Yet data in the Arctic are dispersed and heterogeneous. To make the data more useful, data managers need to better understand the needs of modelers and how they actually access and use data.

This thesis has set out to describe how certain modelers work in detail and to improve understanding of how they assess, prepare, and acquire data for their models. The intention was to identify common, instructive themes or first principles that apply across the models. These themes then suggest data management techniques or requirements for data systems to improve access and use by modelers and generally improve understanding of the Arctic system.

The approach to this research was based in proven social science research methods. The primary method was through the development and analysis of case studies. Case studies have been used extensively to understand business and political processes, but they have rarely been applied as a means to develop data system requirements. And while case studies may not explicitly enumerate system requirements, they do help us understand *how* modelers actually work. This understanding can lead to important insights and challenge some of the assumptions data manages may have made. These insights can then be used to develop propositions for data system improvement as shown in Table 5.2. To refine the propositions and to test their internal and external validity, each case study was presented in a consistent framework and then analyzed independently and in a cross-case synthesis. This question of external validity or applicability beyond an individual case is a central issue in case study research (Gerring, 2007), so each proposition was also tested against broad ethnographic observations described here.

Ethnographic research is challenging, in part, because the observer's presence can obviously influence the observation process and subsequent interpretation. As an ethnographer, I am not an

impartial researcher. As a data manager at NSIDC, I professionally interact with all three of the participants. I developed the initial sea ice product comparison page at NSIDC in response to requests from sea ice researchers like Dr. Holland (Parsons and Duerr, 2005). I led the data management effort for the Cold Land Processes Experiment, the source of much of Dr. Liston's data (Parsons et al., 2004). I met Dr. Rastetter at the Arctic synthesis workshop described in Chapter 1. I was an invited speaker on data and technology issues, and Dr. Rastetter was an outspoken advocate for improved data availability. More importantly, I have been helping lead a broad team of scientists and data managers around the world to create a sustained, interdisciplinary polar data system initially for IPY, but also for polar science more generally. This effort has included a variety of workshops, conference sessions, interoperability experiments, standards assessments, and the creation of new data systems. It is a slow, grass roots, but growing effort, that has learned a lot about interdisciplinary data management and integrated system development (Parsons et al., 2010). The IPY data management experience has greatly informed this study. So while the observations may not be impartial, they are extensive, rich, and supported by professional data management expertise.

With this in mind, we return now to the cross-case comparison. Analyzing each of the case studies independently and in concert has led to many insights and has suggested many principles and practices that were highlighted in chapter 5. These suggestions are valuable in their own right, but it is difficult to test the full validity of any particular assertion with only three case studies. Inevitably more research is necessary, but one important but simple principle emerges: *work on the data first*. Data are more important than systems. Data centers get more return on their investment in reducing user effort by providing consistent, well-described data in the desired format than they do in developing improved data analysis, subsetting, and access tools. This basic principle, focus on the data first, implies immediate action data centers can take to improve modeling efficiency, by providing data in multiple precise formats and harmonizing basic meteorological and hydrological in-situ measurements across multiple stations and networks. In some cases, it may be appropriate to develop specific products for specific communities.

It is clear how data improvement could have improved the efficiency of each modeler be it a through a sea ice data set in self-described format, a consistent way to present air temperature, or a defined split between net and gross mineralization. An example from IPY provides further illustration of how data harmonization increases efficiency. A major multinational, European, IPY project—Developing Arctic Modelling and Observing Capabilities for Long-term Environmental Studies (DAMOCLES)—sought to develop an inexpensive data management system for the project. Early in the project, in response to desires of the modelers, data collectors agreed to provide all their data to the DAMOCLES data system in netCDF-CF. The intent was to reduce data management costs and make more money available for data collection and research. At first, there was some resistance to providing the data in netCDF, but as the data managers worked with the communities and developed tools to help them convert their data, the system began to work well. Very soon after the project began, all the data were available through multiple protocols (ftp, WCS, OpenDAP, KML) and readily useful with many existing tools in the Arctic ocean/ice and climate modeling community—all at a fraction of the original supposed cost. The modeling systems are more efficient and so is the data system because of early agreement and collaboration on a common format (Ø. Godøy, personal communication).

Related to the principle of focusing on data is appropriate credit and accountability for the data. Data citation is a practice that needs to grow and data centers should always provide recommended citations for their holdings. This is a nascent practice but it should be encouraged and further developed. Meanwhile more research is needed on how best to review, assure, and assert the quality of data. Formal studies should be conducted on peer-review schemes, community review through social networking and virtual organizations, and effective means of presenting uncertainty.

A few cautions when considering these results. These case studies all focused on well-established senior scientists. It is possible that an analysis including more junior scientist might have revealed different results. Nevertheless, scientists as a whole tend to be conservative in their methods, and casual observation of more junior modelers suggests they take similar approaches to their data assessment, acquisition, and preparation—i.e. get it all at once and then process it into what they need. Examining a different set of models would also have likely led to different

insights. That said, given the disparity of these three models and the external validation through the IPY and other work, it is unlikely that additional findings would actively refute what was learned from these models. Finally, perhaps most importantly, ignoring the issue of data discovery may have avoided some key issues. Data discovery is clearly a big issue. It could have been such a big issue, though, that it could have overshadowed the details of how modelers assess data. More of these type of studies focused on data discovery could be beneficial.

The first conclusion of this overall analysis is simply that this is an effective study method. By examining *how* modelers work, one learns much more about their real issues and priorities and how they make decisions. One gets a much better understanding of *why* modelers need what they do. Simply asking users what they want is not always revealing or even accurate, whereas understanding how they work reveals the underlying needs. This case study methodology could be a nice additional form of user engagement beyond conventional advisory groups, use cases, and usability studies. Simply conducting the case study provides insight even if each insight is not fully "proven". During the time conducting these studies, I found they often informed my daily decisions as a data manager that ultimately manifested into formal system requirements on some projects and ideas for funding proposals. One of the criticisms of usability studies is that they only identify what is wrong with a system. They do not always provide guidance on new data system approaches. This case study approach can better uncover true needs. When developing a new portal, it would be useful to conduct a few short case studies of how users have worked in the past. This may be criticized as a backward looking approach, but when coupled with the development of formal use cases and agile, iterative development approaches, it is likely to produce a system more consistently useful for the specific community. Indeed these studies would be most effective when targeted around a specific data center and user community.

The more intriguing conclusion is that the best expenditure of limited resources to increase the efficiency of modeling studies is to improve the consistency and flexibility of the data and the documentation rather than enhanced interfaces and analysis tools. There is a demand and need for these tools, but there is greater short-term return (reduction of scientific effort) with improved data, which, in turn, makes it easier to build more effective tools in the long run. It is

likely that this data-first philosophy can improve the data systems that support the overall interdisciplinary, integrative science necessary to understand the complex Arctic system.

Works Cited

ACIA (Arctic Climate Impacts Assessment). 2005. *Impacts of a Warming Arctic, Arctic Climate Impacts Assessment.* Cambridge University Press.

Aitkenhead, JA and WH McDowell. 2000. Soil C: N ratio as a predictor of annual riverine DOC flux at local and global scales. *Global Biogeochemical Cycles* 14 (1): 127-138.

Allison, I, M Beláand, K Alverson, R Bell, D Carlson, K Danell, C Ellis-Evans, *et al.*. 2007. *The Scope of Science for the International Polar Year 2007–2008, WMO/TD–No. 1364.* Geneva: World Meteorological Organization.

Anderson, MP, and WW Woessner. 1992. *Applied Groundwater Modeling: Simulation of Flow and Advective Transport.* San Diego, CA: Academic Press.

Barbero, M. 2006. *Mineralization of dissolved organic carbon (DOC) and nitrogen (DON) in the soil solution of different forested stands in Flanders.* Turin, Italy: Politecnico di Torino. http://www.tesionline.com/intl/thesis.jsp?idt=16741

Barry, R. G., and S. Smith. 2004. Report of the Standing Committee on Data, Information, and Communications. *Frozen Ground vol. 28.*

Bormann, FH, and GE Likens. 1979. *Pattern and process in a forested ecosystem.* New York: Springer-Verlag. 253 pp.

Bourke, RH, and RP Garrett. 1987. Sea ice thickness distribution in the Arctic Ocean. *Cold Regions Science and Technology. 13*:259-280.

Briegleb, P., M Bitz, C Hunke, H Lipscomb, M Holland, L Schramm, and E Moritz. 2004. *NCAR Technical Note NCAR/TN-463+STR.* Boulder, CO: NCAR.

Brooks, PD, DM McKnight, and KE Bencala. 1999. The relationship between soil heterotrophic activity, soil dissolved organic carbon (DOC) leachate, and catchment-scale DOC export in headwater catchments. *Water Resources Research* 35 (6): 1895-1902.

de Bruin, T, R Chen, MA Parsons, and D Carlson. 2009. Freeing data through the Polar information Commons. *Eos Trans. AGU. 90 (52)*:abstract IN34B-302.

Cavalieri, D, C Parkinson, P Gloersen, and HJ Zwally. 1996. *Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data.* Boulder, CO, USA: National Snow and Ice Data Center. http://nsidc.org/data/nsidc-0051.html. Retrieved on 29 Jul. 2010.

Chapin, FS, M Sturm, MC Serreze, JP McFadden, JR Key, AH Lloyd, AD McGuire, *et al.*. 2005. Role of land-surface changes in Arctic summer warming. *Science. 310*:657-660. 10.1126/science.1117368

Chen, R, and MA Parsons. 2010. *Creating a Polar Information Commons.* Presented at Association of American Georgraphers,Washington, DC,18 Apr. 2010.http://dusk.geo.orst.edu/aag_ethics10.html, retrieved 28 Jul. 2010.

Cline, D, R Armstrong, R Davis, K Elder, and G Liston. 2003a. *Clpx-Ground: ISA Snow Depth Transects and Related Measurements.* Ed. MA Parsons and MJ Brodzik. Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/data/nsidc-0175.html. Retrieved on 29 July 2010.

Cline, D, K Elder, B Davis, J Hardy, GE Liston, D Imel, SH Yueh, *et al.*. 2003. Overview of the NASA cold land processes field experiment (CLPX-2002). In CD Kummerow, J Jiang, and S Uratuka (ed.). *Microwave Remote Sensing of the Atmosphere and Environment III.* SPIE. pp. 361-372.

Cline, D, R Armstrong, R Davis, K Elder, and G Liston. 2004. *Clpx-Ground: ISA Snow Pit Measurements.* Ed. MA Parsons and MJ Brodzik. Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/data/nsidc-0176.html. Retrieved on 29 July 2010.

Cline, D and T Carrol. 2004. *CLPX Airborne Gamma Snow and Soil Moisture Surveys.* Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/data/nsidc-0158.html. Retrieved on 29 July 29 2010.

Cline, D, S Yueh, B Chapman, B Stankov, A Gasiewski, D Masters, K Elder, *et al.*. 2009. NASA Cold Land Processes Experiment (CLPX 2002/03): Airborne Remote Sensing. *Journal of Hydrometeorology. 10*:338-346.

Collins, WD, CM Bitz, ML Blackmon, GB Bonan, CS Bretherton, JA Carton, P Chang, *et al.*. 2006. The Community Climate System Model version 3 (CCSM3). *J CLIMATE. 19*:2122-143.

Costello, MJ. 2009. Motivating Online Publication of Data. *BIOSCIENCE. 59*:418-427. 10.1525/bio.2009.59.5.9

Couclelis, H. 2003. The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS. 7*:165-175. doi:10.1111/1467-9671.00138

Doesken, N, and A Judson. 1996. *The Snow Booklet: A Guide to the Science, Climatology, and Measurement of Snow in the United States.* Colorado Climate Center, Colorado State University. 84 pp.

Durner, GM, DC Douglas, RM Nielson, SC Amstrup, TL McDonald, I Stirling, M Mauritzen, *et al.*. 2009. Predicting 21st-century polar bear habitat distribution from global climate models. *ECOL MONOGR. 79*:25-58.

Krupnik, and Jolly (ed.).2002. *The Earth is Faster Now: Indigenous Observations of Arctic Environmental Change.* Fairbanks, AK: Arctic Research Consortium of the United States. 356 pp.

Edwards, PN, S J Jackson, GC Bowker, and CP Knobel. 2007. *Understanding Infrastructure: Dynamics, Tensions, and Design.* Arlington, VA: National Science Foundation. Available at http://hdl.handle.net/2027.42/49353.

Elder, K, D Cline, GE Liston, and R Armstrong. 2009. NASA Cold Land Processes Experiment (CLPX 2002/03): Field Measurements of Snowpack Properties and Soil Moisture. *Journal of Hydrometeorology. 10*:320-29.

Elder, K and A Goodbody. 2004. *Clpx-Ground: ISA Main Meteorological Data.* Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/data/nsidc-0172.html. Retrieved on 29 July 2010.

Fahey, TJ, and JW Hughes. 1994. Fine root dynamics in a northern hardwood forest ecosystem, Hubbard Brook Experimental Forest, NH. *Journal of Ecology. 82*:533-548.

Fetterer, F. 2009. Data Management Best Practices for Sea Ice Observations. In H Eicken, R Gradinger, M Salganek, K Shirasawa, D Perovich, and M Leppäranta (ed.). *Field Techniques for Sea-Ice Research* Fairbanks, AK: University of Alaska Press.

Fetterer, F, and K Knowles. 2002. *Sea Ice Index.* Boulder, CO, USA: National Snow and Ice Data Center. http://nsidc.org/data/g02135.html. Retrieved on 29 Jul. 2010.

Gerring, J. 2007. *Case study research: principles and practices.* New York: Cambridge University Press. x, 265 pp.

Goodale, CL, JD Aber, and WH McDowell. 2000. The long-term effects of disturbance on organic and inorganic nitrogen export in the White Mountains, New Hampshire. *Ecosystems* 3 (5): 433-450.

de Sherbinin A, and RS Chen (ed.).2005. *Global Spatial Data and Information User Workshop: Report of a Workshop.* Palisades, NY: Socioeconomic Data and Applications Center, Center for International Earth Science Information Network, Columbia University.

Herbert, DA, EB Rastetter, L Gough, and GR Shaver. 2004. Species diversity across nutrient gradients: An analysis of resource competition in model ecosystems. *ECOSYSTEMS. 7*:296-310. 10.1007/s10021-003-0233-x

Holland, MM, CM Bitz, and B Tremblay. 2006. Future abrupt reductions in the summer Arctic sea ice. *GEOPHYS RES LETT. 33*:L23503. 10.1029/2006GL028024

Hubbard Brook Ecosystem Study. 2001. *Guide to the Hubbard Brook Experimental Forest and the Hubbard Brook Ecosystem Study.* http://www.hubbardbrook.org/overview/introduction.htm. Accessed 15 Jul. 2010.

ICSU (International Council for Science). 2004. *A Framework for the International Polar Year 2007-2008.*

IPCC. 2000. *IPCC Special Report Emissions Scenarios.* http://www.grida.no/publications/other/ipcc_sr/. Accessed 29 Jul. 2010.

IPCC. Watson RT, and Core Writing Team (ed.).2001. *Climate change 2001: synthesis report: contribution of Working Groups I, II, and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

ISO. 2003. *ISO Standard 14721:2003, Space Data and Information Transfer Systems—A Reference Model for an Open Archival Information System (OAIS),.* International Organization for Standardization.

Key Perspectives Ltd. 2010. *Data dimensions: disciplinary differences in research data sharing, reuse and long term viability.* Edinburgh: Digital Curation Center. http://www.dcc.ac.uk/scarp.

Klump, J, R Bertelmann, J Brase, M Diepenbroek, H Grobe, H Höck, M Lautenschlager, U Schindler, I Sens, and J Wächter. 2006. Data publication in the open access initiative. *Data Science Journal. 5*:79-83. 10.2481/dsj.5.79

LeDrew, E, MA Parsons, and T de Bruin. 2008. Securing the Legacy of IPY. *Earthzine.* . http://www.earthzine.org/2008/03/27/securing-the-legacy-of-ipy/

Lindsay, RW, and J Zhang. 2005. Thinning Arctic Sea ice: Have we passed a tipping point?. *B AM METEOROL SOC. 86*:325-26.

Liston, GE. 2004. *CLPX-Model: Local Analysis and Prediction System: 4-D Atmospheric Analyses.* Boulder, CO: National Snow and Ice Data Center. http://nsidc.org/data/nsidc-0179.html.

Liston, GE, and CA Hiemstra. 2008. A Simple Data Assimilation System for Complex Snow Distributions (SnowAssim). *Journal of Hydrometeorology.* :989-1004. 10.1175/2008JHM871.1

Liston, GE, and JG Winther. 2005. Antarctic surface and subsurface snow and ice melt fluxes. *J CLIMATE. 18*:1469-481.

Liston, GE, and K Elder. 2006a. A distributed snow-evolution modeling system (SnowModel). *J HYDROMETEOROL. 7*:1259-276.

Liston, GE, and K Elder. 2006b. A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet). *Journal of Hydrometeorology. 7*:217-234.

Liston, GE, CA Hiemstra, K Elder, and D Cline. 2008a. Meso-cell study area (MSA) snow distributions for the Cold Land Processes Experiment (CLPX). *Journal of Hydrometeorology. 9*:957-976. 10.1175/2008JHM869.1

Liston, GE, DL Birkenheuer, CA Hiemstra, DW Cline, and K Elder. 2008b. NASA Cold Land Processes Experiment (CLPX 2002/03): Atmospheric Analyses Datasets. *Journal of Hydrometeorology. 9*:952-56.

McGuffie, K, and A Henderson-Sellers. 2005. *A Climate Modelling Primer.* Wiley.

Meek, DW, and JL Hatfield. 1994. Data quality checking for single station meteorological databases. *Agricultural and Forest Meteorology. 69*:85-109. doi: DOI: 10.1016/0168-1923(94)90083-3

Meier, WN, ML VanWoert, and C Bertoia. 2001. Evaluation of operational SSM/I algorithms. *Annals of Glaciology. 33*:102-08.

Mernild, SH, GE Liston, B Hasholt, and NT Knudsen. 2006. Snow distribution and melt modeling for Mittivakkat Glacier, Ammassalik Island, southeast Greenland. *J HYDROMETEOROL. 7*:808-824.

NASA SEEDS. 2003. *Strategic Evolution of Earth Science Enterprise Data Systems (SEEDS) Formulation Team Final Recommendations Report.* July. http://lennier.gsfc.nasa.gov/seeds/FinRec.htm. Accessed 15 Jan. 2004.

Neff, JC, SE Hobbie, and PM Vitousek. 2000. Nutrient and mineralogical control on dissolved organic C, N and P fluxes and stoichiometry in hawaiian soils. *Biogeochemistry* 51 (3): 283-302.

Nelson, B. 2009. Data sharing: Empty archives. *Nature. 461*:160-63. http://www.nature.com/news/2009/090909/full/461160a.html

NRC (National Research Council). 2006. *Toward an Integrated Arctic Observing Network.* Washington, DC: National Academies Press.

NRC (National Research Council). 2007. *Environmental Data Management at NOAA: Archiving, Stewardship, and Access.* Washington, DC: National Academies Press. 116 pp.

OMB (Office of Management and Budget). 2002. *Circular A-16, revised: Coordination of Geographic Information, and Related Spatial Data Activities 19 August 2002.*

Overpeck, JT, M Sturm, JA Francis, DK Perovich, and MC Serreze. 2005. Arctic System on Trajectory to New, Seasonally Ice-Free State. *Eos, Transactions of the American Geophysical Union. 86*.

Parsons, Mark A., Ruth Duerr, and Jean-Bernard Minster. 2010. Data citation and peer-review. *Eos, Trans. AGU 91* (34): 297-298.

Parsons, MA. 2006. International Polar Year Data Management Workshop, 3-4 March 2006, Cambridge, UK. *Glaciological Data Series. GD-33*. http://nsidc.org/pubs/gd/Glaciological_Data_33.pdf

Parsons, MA, and BE Wilson. 2007. User-driven design of a data system for the International Polar Year. *Eos, Transactions of the American Geophysical Union. 88*.

Parsons, MA, and R Duerr. 2005. Designating user communities for scientific data: challenges and solutions. *Data Science Journal. 4*:31-38.

Parsons, MA, MJ Brodzik, and NJ Rutter. 2004. Data management for the cold land processes experiment: improving hydrological science. *HYDROL PROCESS. 18*:3637-653. 10.1002/hyp.5801

Parsons, MA, T de Bruin, S Tomlinson, H Campbell, Ø Godøy, J LeClert, and IPY Data Policy and Management SubCommittee. 2010 (in press). The State of Polar Data. In D Hik, and I Krupnik (ed.). *Understanding Earth's Polar Challenges: International Polar Year 2007-2008* Springer.

Pressman, JL, and A Wildavsky. 1973. *Implementation: How great expectations in Washington are dashed in Oakland.* Berkeley, CA: University of California Press.

Rastetter, EB, and GR Shaver. 1992. A model of multiple-element limitation for acclimating vegetation. *Ecology. 73*:1157-174.

Rastetter, EB, GI Agren, and GR Shaver. 1997. Responses of N-limited ecosystems to increased CO2: A balanced-nutrition, coupled-element-cycles model. *ECOL APPL. 7*:444-460.

Rastetter, EB, PM Vitousek, C Field, GR Shaver, D Herbert, and GI Agren. 2001. Resource optimization and symbiotic nitrogen fixation. *ECOSYSTEMS. 4*:369-388.

Rastetter, EB, SS Perakis, GR Shaver, and GI Agren. 2005. Terrestrial C sequestration at elevated-CO2 and temperature: The role of dissolved organic N loss. *ECOL APPL. 15*:71-86.

Raymond, ES. 2004. *The Art of Unix Programming.* Addison-Wesley Professional.

Rayner, NA, DE Parker, EB Horton, CK Folland, LV Alexander, DP Rowell, EC Kent, and A Kaplan. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century . *J. Geophys. Res.. 108*. 10.1029/2002JD002670

SEARCH (Study of Environmental Change). 2005. *Study of Environmental Change: Plans for Implementation During the International Polar Year and Beyond.* Fairbanks, AK: Arctic Research Consortium of the United States. 104 pp.

Serreze, MC, and RG Barry. 2005. *The Arctic climate system.* New York: Cambridge University Press.

Serreze, MC, JE Walsh, FS Chapin, T Osterkamp, M Dyurgerov, V Romanovsky, WC Oechel, J Morison, T Zhang, and RG Barry. 2000. Observational evidence of recent change in the northern high-latitude environment. *CLIMATIC CHANGE. 46*:159-207.

Sturm, M, C Racine, and K Tape. 2001. Climate change - Increasing shrub abundance in the Arctic. *Nature. 411*:546-47.

Sturm, M, J Holmgren, and GE Liston. 1995. A seasonal snow cover classification-system for local to global applications. *J CLIMATE. 8*:1261-283.

Whittaker, RH, GE Likens, FH Bormann, JS Easton, and TG Siccama. 1979. The Hubbard Brook ecosystem study: forest nutrient cycling and element behavior. *Ecology. 60*:203-220.

Yin, RK. 2003. *Case study research: design and methods.* Sage Publications. xvi, 181 pp.

Appendix A:
ACRONYM LIST

AAG:	Association of American Geographers
AGU:	American Geophysical Union
AON:	Arctic Observing Network
CCSM:	Community Climate System Model version 3
CLPX:	Cold Land Processes Experiment
CMIP:	Coupled Model Intercomparison Project
CSIM:	Community Sea Ice Model
CSIM:	Community Sea Ice Model version 5
DAMOCLES:	Developing Arctic Modelling and Observing Capabilities for Long-term Environmental Studies
DIN:	Dissolved Inorganic Nitrogen
DON:	Dissolved Organic Nitrogen
FTP:	file transfer protocol
GCM:	Global Climate Model
HadISST:	Hadley Centre Global Sea Ice and Sea Surface Temperature
IOP:	Intensive Observation Period
IPCC:	Intergovernmental Panel on Climate Change
ISA:	Intensive Study Area
LAPS:	Local Analysis and Prediction System
MEL:	Multiple Element Model
MSA:	Mesoscale Study Area
NASA:	National Aeronautics and Space Administration
NCAR:	National Center for Atmospheric Research
NED:	National Elevation Data Set
netCDF-CF:	network Common Data Format with Climate Forecast Extensions
NOAA:	National Oceanographic and Atmospheric Administration
NRC:	National Research Council
NSF:	National Science Foundation
NSIDC:	National Snow and Ice Data Center
OpenDAP:	Open-source Project for a Network Data Access Protocol
PCMDI:	Program for Climate Model Diagnosis and Intercomparison
RAWS:	Remote Automated Weather Station
SAON:	Sustained Arctic Observing Network
SEARCH:	Study of Arctic Environmental Change
SNOTEL:	Snow Telemetry
SWE:	snow water equivalent
USGS:	United States Geological Survey
WCRP:	World Climate Research Programme
WCS:	Web Coverage Service
WDC:	World Data Center

Appendix B:
Case Study Interview Protocol

This is a general protocol to guide the interview portion of the three case studies in this thesis. It is intentionally open and not overly detailed to allow the participants to guide the discussion in ways that are relevant to their situation. Nevertheless, to aid analysis, it is necessary to have a certain degree of consistency across the three studies and to ensure certain topics are addressed. This protocol provides an approach to ensure that consistency and completeness.

1. Prior to the first interview.
    a. Obtain agreement from the modeler to participate. Provide background and rationale on my study to the participant. This can be an e-mail or phone call or could include full details of the project if the participant desires. It will include a description of the general approach and specifics of how much time and effort will be involved. Clearly state that the participant will be named and acknowledged in the study in accordance with their wishes and will have the opportunity to review and accept how they are represented before any publication.
    b. Agree with the participant on a particular application of their model to study. The research and application behind a recent paper, for example.
    c. Ask the participant for a list of data used in the study.
2. First interview (~2 hours)
    a. General background
        i. Have them describe the science questions they were trying to address with the particular model application and more generally.
        ii. Have the participant outline their overall research approach for the model application in question, highlighting when they need to use data. Did they follow a formal protocol (e.g., Anderson and Woessner 1992)?
    b. Assessment
        i. Develop/refine a list of data used by the modeler in the agreed application. This will likely require some targeted questions around the approach that they described above. This will be a baseline to refer to in detailed questioning.

ii. Create a table indicating the purpose of each data collection (e.g., evaluation, assimilation) and the primary criteria the modeler used to assess the applicability of the data (e.g., spatial or temporal scale or coverage, accuracy, format, relation to tools). Explore what criteria were most important and what compromises in the data the modeler had to accept or work around. What defines data "consistency"? How important is it?

iii. Develop a similar table for data that are desirable but not available.

iv. Referring to the table(s). Have the participant describe how they assess various data (e.g., documentation, scientific articles, talk to experts at a data center or the data provider, reputation of provider). What was missing? What would have made assessment easier? Tools, info presentation, greater knowledge by data center, attribution, etc.

v. Explore how the application of the data in the study influences the assessment. Is there a relationship between the presentation and format of the final product and the input data?

vi. Probe assumptions. Why is the modeler making certain decisions? What data assessment criteria are taken for granted? Use specific data sets as examples: Why did you use a particular data set? What was good about it? What was lacking?

vii. Use concept maps, flow diagrams, or other tools to get the participant to illustrate their overall assessment process. Consider the actual and idealized situation. Start to sketch out causal linkages (cf. Gerring 2007). Employ white boards or lots of scratch paper. This can continually be developed in the subsequent sections of the interview.

c. Acquisition

i. Add to the data set table how each data set was acquired. Describe source, data transfer method.

ii. Where do the data need to be? Modeler's workstation? Supercomputer? Available though standard protocols (e.g. OpenDAP, OGC)?

iii. What problems restricted acquisition? Any disconnect between access and discovery? Data release, timeliness issues, formats or media? Describe a more ideal process.

iv. Did you subset or resample the data in any way? Would you have liked to?

d. Preparation

i. Have the participant describe the general data preparation process for each class of data however defined (e.g. assimilation or evaluation). Is there a need to develop some level of consistency?

ii. Delve into appropriate particulars. Are there issues of format, down- or upscaling, cross data set integration, data interpolation, etc?

3. Prior to second interview

a. Send participant use case diagrams, concept maps, or other figures; the complete data set table; and brief descriptions of initial conclusions.

4. Second interview (30-60 min)

a. Follow up with any questions necessary to remove gaps in understanding or to remove ambiguity from first interview

b. Review and revise figures and text provided earlier.

Appendix C:
Detailed Summary Tables of Data Used by the Modelers

Table C-1. Data used in Holland et al. (2006)

Stage	Data Collection/Data Set	Application	Data Source	Other data considered	Evaluation Criteria	Data Access Method	Pre-processing (remote)	Data Prep.
Data0	inputs to CMIP3/CCSM models	initalization	n/a (production runs done by others)	n/a	worked with the model	n/a	no	n/a
Data1	CCSM ensemble runs	analysis	local	PCMDI model runs	convenience--available locally	internal	no	no
Data1	WCRP CMIP3 multi-model dataset (15 models)	analysis	PCMDI/Earth System Grid: https://esg.llnl.gov:8443/home/publicHomePage.do	CCSM	convenience--readily available (still only used when requested by reveiewer) convenience--in standard form at PCMDI	ftp	no	regrid
Data2	Hadley Centre Global Sea Ice and Sea Surface Temperature (HadISST) (Rayner et al., 2003)	initial assesment	NCAR Research Data Archive http://dss.ucar.edu/datasets/ds277.3/	SSM/I	parameter to match model output long, current time series format (netCDF) grid validity--discussion with colleagues	ftp	no	regrid
Data2	Sea Ice Index (Fetterer, and Knowles, 2002)	initial assesment	NSIDC: http://nsidc.org/data/g02135.html	none	quick look	http	select period	no
Data2	Ice thickness (Bourke, and Garrett, 1987)	initial assesment	Bourke and Garnett 1986	none	all that's available, hasn't considered Rothrok	literature	n/a	no
Data3	Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I Passive Microwave Data (Cavalieri et al., 1996)	full assessment	NSIDC: http://nsidc.org/data/nsidc-0051.html	HadISST	parameter to match model output long, current time series accuracy	ftp	no	regrid to model grid; reformat to netCDF
Data3	ice thickness (Bourke, and Garrett, 1987)	full assessment	Bourke and Garnett 1986	none	all that's available, hasn't considered rothrok	literature	n/a	no
Data4	Special Report on Emissions Scenarios forcings	forcing	standard simulations	n/a	n/a, it's the benchmark	n/a	n/a	no

Table C.2 Data used in Liston et al. (2008)

Stage	Data Collection/Data Set	Application	Data source	Other data considered	Evaluation criteria	Data access method	Pre-processing (remote)	Data prep.
0	LAPS inputs	input to LAPS						
data1a	CLPX main met stations(precipitation, wind speed and direction, air temperature, and relative humidity)	MicroMet forcing and assimulation	cited as Elder et al 2009; as calculated from http://nsidc.org/data/nsidc-0172.html	other local towers or regional networks	location (ine MSA) temporal resolution (hourly or better) accessible on the internet (convenient) personal knowledge of terrain and measurement quality	ftp	site selection	MicroMet preprocessing including missing value ID, basic QC (Meek and Hatfield 1994), complete missing values (Liston and Elder 2006). Wind assumed missing under canopy
data1a	Fraser Experimental Forest Met stations	MicroMet forcing and assimulation	From Kelly directly or from http://www.fs.fed.us/rm/fraser/data/index.shtml?	other local towers or regional networks	location (ine MSA) temporal resolution (hourly or better) accessible on the internet (convenient) personal knowledge of terrain and measurement quality	ftp	site selection	MicroMet preprocessing including missing value ID, basic QC (Meek and Hatfield 1994), complete missing values (Liston and Elder 2006). Wind assumed missing under canopy
data1a	SNOTEL temperatures	MicroMet forcing and assimulation	http://www.wcc.nrcs.usda.gov/snow/snotel-temp-data.html	other local towers or regional networks	location (ine MSA) temporal resolution (hourly or better) accessible on the internet (convenient) personal knowledge of terrain and measurement quality	ftp	site selection	MicroMet preprocessing including missing value ID, basic QC (Meek and Hatfield 1994), complete missing values (Liston and Elder 2006). Wind assumed missing under canopy
data1a	CLPX Flux Tower	MicroMet forcing and assimulation	internet	other local towers or regional networks	location (ine MSA) temporal resolution (hourly or better) accessible on the internet (convenient) personal knowledge of terrain and measurement quality	ftp	site selection	MicroMet preprocessing including missing value ID, basic QC (Meek and Hatfield 1994), complete missing values (Liston and Elder 2006). Wind assumed missing under canopy

data1a	RAWS Met	MicroMet forcing and assimulation	internet	other local towers or regional networks	location (ine MSA) temporal resolution (hourly or better) accessible on the internet (convenient) personal knowledge of terrain and measurement quality	ftp	site selection	MicroMet preprocessing including missing value ID, basic QC (Meek and Hatfield 1994), complete missing values (Liston and Elder 2006). Wind assumed missing under canopy
data1a	DRI met	MicroMet forcing and assimulation	internet	other local towers or regional networks	location (ine MSA) temporal resolution (hourly or better) accessible on the internet (convenient) personal knowledge of terrain and measurement quality	ftp	site selection	MicroMet preprocessing including missing value ID, basic QC (Meek and Hatfield 1994), complete missing values (Liston and Elder 2006). Wind assumed missing under canopy
data1b	Local Analysis and Prediction System (LAPS) analyses	MicroMet forcing and assimulation	Dan Birkenheuer acknowledged (Liston et al 2008)	none	consistency of inputs, especially precipitation	ftp	spatial and temporal subsettin g	convert gridded values to point values
data2	USGS National Elevation Dataset	MicroMet and SnowModel forcing	http://ned.usgs.gov/	none	high resolution coverage consistency reuse	ftp	spatial subsettin g	reprojected and gridded from GIS format
data2	National Land Cover Data	MicroMet and SnowModel forcing		Vogelmann et al 2001	none	high resoluti on coverag e consist ency reuse	ftp	spatial subsetting
data3	Gravimetric Soil moisture	Gamma calibration/c orrection	personally from Elder	n/a	created for project	individu al	no	used by Cline and Elder to calculate Gamma SWE
data3	SWE from Gamma data	SnowModel assimilation	Derived from data at NSIDC by Elder (Cline et al. 2004)	n/a	created for project	individu al		Recalculated NP IOP3 data using measured soil moisture. Multilply Each North Park GAMMA value by the ratio of the ISA average to the GAMMA average for Each IOP. Interpolate to simulation grid

Table C.3 Data Used in Rastetter et al. (2005)

Stage	Data Collection/Data Set	Application	Data source	Other data considered	Evaluation criteria	Data access method	Pre-processing (remote)	Data prep.
Data1	Hubbard brook etc.	paramaterization, forcing	stocks and fluxs from sources described in Rastetter et al. (2001) table 1 then used to calculate parameters in table 2. Mostly peer-reviewed refs with some pers. Communication.	HJ Andrews. Arctic LTER	Comprehesiveness--fine root dynamics! Authoratative/citable	transcription and interpretation	none	unit conversion and other calculations
Data1	C:N ratio	parameterization	Goodale, et al. 2001	none	fit with Hubbard Brook data	transcription and interpretation	none	little or none
Data2	Microbial respiration	parameterization	calculated modification to values in Rastetetter 2001 to maintain assumtion of steady state	n/a	inherent in how N processes are described in the model	n/a	none	little or none
Data2	Gross N mineralization	parameterization	calculated modification to values in Rastetetter 2001 to maintain assumtion of steady state	n/a	inherent in how N processes are described in the model	n/a	none	little or none
Data2	N immobilization	parameterization	calculated modification to values in Rastetetter 2001 to maintain assumtion of steady state	n/a	inherent in how N processes are described in the model	n/a	none	little or none
Data3	2x CO2 and 4°C temperature increase	forcing	IPCC 2001 (for New England)	none	benchmark	transcription and interpretation	none	little or none
Data3	M1: constant DOC loss	parameterization	baseline assumption	none	none	n/a	none	little or none
Data3	M2 constatnt: proportional to organic matter in the soil	alternate parameterization	based on values from Neff et al. 2000	none	authoratative	transcription and interpretation	none	little or none
Data3	M3 constant: proportional to C:N ratio	alternate parameterization	based on values from Aitkenhead and McDowell 2000	none	authoratative	transcription and interpretation	none	little or none

Data3	M4 constant: proportional to microbial respiration	alternate parameterization		based on values from Brooks et al 1999	none	authoratative	transcription and interpretation	none
Data4	Calculated plant and soil C and N stocks	assessment	Derived from original MEL (Rastetter, et al. 1992)	n/a	inherent in model	n/a	none	little or none