Evaluating the effect of accordance measure choice on the selection of
ground surface temperature models in a sub-Arctic environment

by

Luis Padilla-Ramirez

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Master of Science

in

Geography, Data Science Specialization

Carleton University
Ottawa, Ontario

© 2020
Luis Padilla-Ramirez

ABSTRACT

Thawing of permafrost has implications on global climate, but collecting ground observations to monitor its evolution is resource-intensive and consequently, observations are sparse. Computational models are powerful tools used to fill gaps in information required to inform sound policy decision-making but in the study of permafrost, such models are often validated using simple ad-hoc comparisons against observations. In this study, six ground-surface temperature (GST) models are compared against observations and their performance ranked by multiple mathematical accordence measures. A small variation in model ranks is present between accordence measures. Subdivision of the observations temporally by season and spatially by ground type results in greater rank variability, and reveals biases in the models and accordence measures. These results demonstrate the importance of a well-defined objective to guide the selection of accordence measures and observation data subsets to be used in selecting a “best model” on which policy decisions can be based.

ACKNOWLEDGEMENTS

I would first and foremost like to acknowledge my thesis supervisor, Dr. Stephan Gruber for his guidance and patience during the course of this research. His concise and insightful input made this work possible. I would also like to thank Dr. Jason Nielsen, who served as the other member of my Thesis Advisory Committee. His advice helped delineate the direction and mathematical details of this thesis. I am indebted to Nick Brown, Bin Cao, and Emilie Stewart-Jones for being so generous with their time in teaching me the ins and outs of the data and processes on which this work is based.

I would also like to acknowledge my wife Courteney, whose support and encouragement has been relentless and unwavering through some very trying times.

Table of Contents

Abstract	ii
Acknowledgements	iii
Table of Contents	viii
List of Tables	ix
List of Symbols Used	xiii
List of Abbreviations, Acronyms, and Initials	xiii
List of Figures	xiii
1 Introduction	1
1.1 Context	1
1.1.1 Permafrost	1
1.1.2 Ground surface temperature	3
1.1.3 Model validation	4
1.2 Goals, aims and objectives	5
1.3 Methodology	10
1.4 Scope	11

1.5	Structure	12
2	Background	13
2.1	Modelling cold environments	13
2.1.1	Introduction	13
2.1.2	Driving data	14
2.1.3	Permafrost modelling	15
2.2	Model validity and selection	16
2.2.1	Modelling objectives	17
2.2.2	Current practices in model validation	17
2.3	Accordance	20
2.3.1	Sum-of-squares-based and sum-of-absolute-error-based measures	21
2.3.2	Dimensioned vs. dimensionless accordance measures	22
2.3.3	Purpose and ranges of accordance measures	25
2.3.4	RMSE and MAE	25
2.3.5	Coefficient of determination (r^2)	26
2.3.6	Nash-Sutcliffe efficiency (NSE)	26
2.3.7	Modified coefficient of efficiency (E_1)	27
2.3.8	Willmott's index of agreement (d)	28
2.3.9	Willmott's modified index of agreement (d_1)	28
2.3.10	Willmott's refined index of agreement (d_r)	29
3	Data and Models	31
3.1	Introduction	31
3.2	Observations	31
3.3	Reanalysis data	35

3.3.1	GlobSim	35
3.3.2	ERA5	35
3.3.3	Japanese 55-year reanalysis (JRA-55)	35
3.3.4	Use in model validation	36
3.4	Models	37
3.4.1	The GEOtop model	37
3.4.2	GEOtop model configuration parameters	38
3.4.3	Parameters representing ground types	39
3.4.4	Simple models	39
3.4.5	Resulting models	39
4	Methods	41
4.1	Introduction	41
4.2	Model runs	41
4.3	Model ranking	42
4.4	Accordance measure selection	42
4.5	Confidence intervals	49
5	Experimentation	52
5.1	Outline	52
5.2	Rankings by annual accordance	53
5.3	Rankings by seasonal accordance	57
5.4	Rankings by seasonal and ground type accordance	59
6	Results	61
6.1	Model evaluation based on annual accordance values	61
6.2	Model evaluation based on seasonal accordance values	64

6.3	Model evaluation based on seasonal performance over different ground types	71
7	Interpretation	77
7.1	Objective 1: The effect of using different measures of accordance in evaluating model competence in GST prediction using annual data	77
7.2	Objective 2: The effect of temporal subdivision of datasets into seasons on evaluations of model competence in GST prediction	79
7.3	Objective 3: The effect of the consideration of ground characteristics on evaluations of model competence for GST prediction using seasonal datasets	82
8	Discussion	84
8.1	Rank variations and their causes	84
8.2	The effect of more specific data subsets on rank variations	85
8.3	Reproducibility	86
8.4	The value of interpretability	86
8.5	Implications of findings on permafrost research	87
8.6	Selecting an accordance measure for GST	88
8.7	Assumptions and further uncertainties	89
9	Conclusion	91
9.1	The effect of different measures of accordance on determinations of model competence in the prediction of GST using annual data	91
9.2	The effect of seasonality on rankings of model accordance across accordance measures	92
9.3	The effects of spatial subdivision by grouping ground types on seasonal measures of accordance	93
9.4	Final thoughts	94

References	96
Appendix A Table of permafrost model validation procedures from nine publi- cations	104
Appendix B Terrain type parameters	110
Appendix C Annual data tables and histograms	113
Appendix D Seasonal data accordances and ranks	123
Appendix E Spatial data accordances and ranks	124

List of Tables

2.1	Summarized model evaluation procedures of nine permafrost modelling papers.	19
2.2	Dimensioned measures of accordance.	23
2.3	Dimensionless measures of accordance.	24
3.1	Seasonal start dates and number of complete seasons of point observations data available.	33
3.2	Number of observation points and their ground types.	34
3.3	Meteorological variables required by GEOtop, and their units/formats. . . .	37
3.4	Model names, referencing numbers, and description for models used in ex- perimentation.	40
4.1	Accordance measure groupings based on model rankings and common basis.	43
4.2	Model annual accordance ranks and means as calculated by different accor- dance measures.	43
4.3	Summer ranks of models.	45
4.4	Fall ranks of models.	45
4.5	Winter ranks of models.	46
4.6	Spring ranks of models.	46
5.1	Sample model rankings by accordance function.	57

5.2	Sample model/season rankings by accordance function.	59
6.1	Model annual accordance means and ranks as calculated by different accordance measures.	62
6.2	Mean of annual accordance means vs mean of seasonal accordance means. .	64
6.3	Rank stability table for models over all seasons.	65
6.4	Overall model/season ranks.	66
6.5	Summer ranks of models based on terrain and accordance measure.	74
6.6	Fall ranks of models based on terrain and accordance measure.	74
6.7	Winter ranks of models based on terrain and accordance measure.	75
6.8	Spring ranks of models based on terrain and accordance measure.	76
A.1	List of validation procedures used in selected Permafrost publications . . .	109
B.1	GEOtop soil parameters for different subsurface materials.	111
B.2	GEOtop soil, snow and vegetation parameters for different ground types. . .	112
C.1	Annual accordance means, ranks, and 95% confidence intervals for six models using $RMSE$	113
C.2	Annual accordance means, ranks, and 95% confidence intervals for six models using d_1	114
C.3	Annual accordance means, ranks, and 95% confidence intervals for six models using E_1	114
D.1	Mean annual accordance vs mean of seasonal accords.	123

List of Symbols Used

D	Difference between P and O values ($D = P - O$)
O	Set of all observations in data sample
P	Set of all model-produced predictions
M	Set of all models
F	Set of all accordance functions
Y	Set of all points for which complete observations are available
V	Set of all ground types classifications present in O (10 identified)
o_y	Observations at individual point y
$o_{y,t,x}$	Observations at individual point y during time interval type t , where x indicates an individual time interval and a plain t denotes all intervals of type t within o_y . Valid interval types for t are A (full year), Su , Fa , Wi , Sp for summer, fall, winter and spring respectively
$o_{y,t,x}^v$	Observation of individual point y at interval t , where y is a point of ground type group v
$p_{y,t,m}$	Predictions at point y subdivided by time period indicated by t as generated with individual model m
$f(o_{y,t}, p_{y,t,m})$	Accordance function f applied on observation $o_{y,t}$ and model prediction $p_{y,t,m}$
τ	Set of all accordance values
$\tau_{p_{y,t,m}}^f$	Accordance value of function f on model prediction $p_{y,t,m}$ against the corresponding observations at $o_{y,t}$
μ	The mean of the population from which O is sampled
$\bar{x}$	Mean accordance of a model over all individual points in sample O .
$Z_f(A)$	Ranking function of accordance function f for all values in A , where A is a set of numbers. Function returns a list of ranks
$\zeta_{y,t,m}^f$	Individual rank of model m within point $o_{y,t}$ using accordance function f

List of Abbreviations, Acronyms, and Initials

d or d_2	Index of agreement from Willmott (1981)
d_1	Modified index of agreement from Willmott et al. (2012)
d_r	Refined index of agreement from Willmott et al. (2012)
E	See NSE
E_1	Modified coefficient of efficiency from Legates and McCabe (1999)
ECMWF	European Centre for Medium-range Weather Forecasts
ERA5	ECMWF ReAnalysis, fifth major reanalysis product
ERA-Interim	ECMWF ReAnalysis, Interim
GST	Ground surface temperature
ILAMB	International LAnd Model Benchmarking
JMA	Japanese Meteorological Agency
JRA-25	Japanese 25-year ReAnalysis
JRA-55	Japanese 55-year ReAnalysis
MAD	Mean-absolute deviation
MAE	Mean-absolute error
MAGST	Mean annual ground-surface temperature
MBE	Mean-bias error
MERRA-2	Modern-Era Retrospective analysis for Research and Applications, Version 2
MSE	Mean-square error
NCEP	National Centers for Environmental Prediction
NCAR	National Center for Atmospheric Research
NSE	Nash-Sutcliffe efficiency from Nash and Sutcliffe (1970)
r^2	Coefficient of determination from Bennett et al. (2013)
RMSE	Root-mean-square error

List of Figures

3.1	Graph of the distribution of days present in incomplete year timeseries. . . .	33
4.1	Graph of ground surface temperature at point Bushloggers_BS03 in winter 2015, demonstrating a temperature pattern that results in extreme negative <i>NSE</i> values.	48
4.2	Graph of ground surface temperature at point NGO-DD-2035_ST03 in winter 2015 demonstrating a temperature pattern that results in large negative <i>NSE</i> values.	49
5.1	Data flow diagram symbology	52
5.2	Diagram of experiments conducted	53
5.3	Diagram representing subdivision of complete point timeseries into annual timeseries	54
5.4	Diagram representing comparison of observation and prediction timeseries using an accordance measure	55
5.5	Diagram representing the process of calculating mean accordance of all τ values for each model	56
5.6	Diagram representing the ranking of mean accordances of the predictions made by six models against the entire observational dataset	56

5.7	Diagram representing subdivision of complete point timeseries into seasonal timeseries	58
5.8	Diagram representing subdivision of observation points into ground type groupings	60
6.1	95% confidence intervals of all models on annual data per accordance measure.	62
6.2	Plausible model annual ranks for each model based on 95% confidence intervals of all models per accordance measure.	63
6.3	Plausible model ranks for summer based on 95% confidence intervals.	67
6.4	Plausible model ranks for fall based on 95% confidence intervals.	68
6.5	Plausible model ranks for winter based on 95% confidence intervals.	69
6.6	Plausible model ranks for spring based on 95% confidence intervals.	70
6.7	Graph of the mean accordances of different ground types on different models by season as output by $RMSE$	71
6.8	Graph of the mean accordances of different ground types on different models by season as output by d_1	72
6.9	Graph of the mean accordances of different ground types on different models by season as output by E_1	72
6.10	Plausible model ranks for Sedges and Grass in spring at 95%, 80% and 70% confidence intervals.	75
B.1	The 10 ground types used to group observations.	110
B.2	Soil profiles of different terrain types.	111
C.1	Histograms of r^2 over six models using annual datasets.	115
C.2	Histograms of MAE over six models using annual datasets.	116
C.3	Histograms of $RMSE$ over six models using annual datasets.	117

C.4	Histograms of NSE over six models using annual datasets.	118
C.5	Histograms of d over six models using annual datasets.	119
C.6	Histograms of d_1 over six models using annual datasets.	120
C.7	Histograms of d_r over six models using annual datasets.	121
C.8	Histograms of E_1 over six models using annual datasets.	122
E.1	Plausible model ranks for each model in summer based on accordance measures $RMSE$, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), and Snowdrifts (SD).	125
E.1	Plausible model ranks for each model in summer based on accordance measures $RMSE$, d_1 and E_1 on ground types Sedge fen (SF), Sedges and grass (SG), Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).	126
E.2	Plausible model ranks for each model in fall based on accordance measures $RMSE$, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), Snowdrifts (SD), Sedge fen (SF), Sedges and grass (SG).	127
E.2	Plausible model ranks for each model in fall based on accordance measures $RMSE$, d_1 and E_1 on ground types Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).	128
E.3	Plausible model ranks for each model in winter based on accordance measures $RMSE$, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), Snowdrifts (SD), Sedge fen (SF), and Sedges and grass (SG).	129

E.3	Plausible model ranks for each model in winter based on accordance measures $RMSE$, d_1 and E_1 on ground types Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).	130
E.4	Plausible model ranks for each model in spring based on accordance measures $RMSE$, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), Snowdrifts (SD), Sedge fen (SF), and Sedges and grass (SG).	131
E.4	Plausible model ranks for each model in spring based on accordance measures $RMSE$, d_1 and E_1 on ground types Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).	132

Chapter 1

Introduction

1.1 Context

This thesis presents a study of the influence model error statistic selection has on model validation. Models are evaluated on their ability to predict Ground Surface Temperature (GST) at geographic point locations where time series of ground temperature measurements near the surface exist. Subdivision of time series data into years and seasons, and grouping by ground type allows for the assessment of model performance and accordance measure consistency based on these dimensions, and the assignment of a confidence level to model results based on the environmental configuration being simulated. These steps extend previous practice of assessing model performance relative to other models at a single point without consideration of seasonality.

1.1.1 Permafrost

Permafrost is defined as ground (soil or rock and included ice and organic material) that remains at or below 0°C for at least two consecutive years (Van Everdingen, 2005). Estimates indicate that 19.1×10^6 to 24.7×10^6 km² (14% to 19%) of the world's exposed land surface

can be classified as permafrost region, meaning regions underlain by varying spatial proportions of permafrost. The permafrost area, the estimate of exposed land surface area actually underlain by permafrost accounting for discontinuity is 13.2×10^6 to 18.1×10^6 km² (9% to 14%) of the exposed land surface (Gruber, 2012). The active layer is ground that is subject to annual thawing and freezing underlain by permafrost (cf. Van Everdingen, 2005). Active-layer thickness increase, rising mean annual ground temperatures, and decreases in the duration of seasonal ground freezing have been observed on the Tibetan Plateau, in North America, and indeed almost the entire globe has experienced surface warming (Smith et al., 2010; Yang et al., 2010; Stocker et al., 2013) factors that undermine permafrost stability. The stability of permafrost as a ground material to support infrastructure is important, as ground settlement from permafrost thaw can lead to extensive damage to infrastructure (United States Arctic Research Commission Permafrost Task Force, 2003; de Grandpré et al., 2012; Hong et al., 2014). Even though the northern permafrost zone accounts for only 15% of the global soil area, new estimates indicate these surface soils up to 3 m in depth could contain up to 37% of the surface carbon pool on earth. Under the current warming trajectory, estimates indicate that 5% to 15% of the terrestrial permafrost carbon pool is vulnerable to release in the form of greenhouse gases this century, with the majority in the form of CO₂, further contributing to climate change (Schuur et al., 2015). The infrastructure and ecological instability resulting from the thawing of the ground may have very severe outcomes.

Permafrost is primarily a subsurface phenomenon that is not easily observable and therefore, modelling remains the most powerful tool available to determine its spatial extent and rate of degradation. Models are simplified representations of real phenomena that require assumptions, inferences and input parameters, and whose accuracy cannot be definitively verified (Oreskes et al., 1994). To address stakeholder needs, policymakers and researchers will demand rigorous methods of model evaluation if they are to base their decisions on model results. Progress in earning and cultivating trust in models is achieved through sys-

tematic evaluation of their results, identification of their limitations, and improvement of model mechanics in successive iterations.

1.1.2 Ground surface temperature

Ground surface temperature (GST) is measured about 10 cm below the atmosphere-lithosphere interface and is affected by factors including micro-topography, surface cover such as vegetation and snow, local hydrology and lithology (Luo et al., 2018). Solar radiation also affects GST, particularly in mountainous regions where varied slope angles mean varied amounts of exposure (Gubler et al., 2013). The potential variation of these factors means that GST can vary even over short distances. It is important that any GST model validation method consider this spatial variability in surface configuration, which is reflected in the resulting simulated time-series. Within a permafrost system, GST can serve as the upper boundary condition for models, affecting the thermal state of the active layer and permafrost (Chen et al., 2018). This role as both an input and an indicator of active-layer and permafrost temperatures, along with the relative simplicity of its measurement and availability of observations makes GST an attractive metric for study.

Surface air temperature, typically measured about 2 m above the ground surface, and land surface temperature, which corresponds to the temperature of the top of the surface cover are two other kinds of surface temperature measurements related to the thermal regime of an area underlain by permafrost which, though connected to GST through the complex heat exchange processes at the ground surface and within the atmosphere, differ greatly (Luo et al., 2018). Use of surface air temperature for predicting permafrost presence requires caution as other key factors, such as the snow cover and vegetation at local scales are not accounted for in this measure. Land surface temperature does not completely account for the influence of the soil/litter layer and surface characteristics, or the buffering effects of

vegetation, snow cover, and upper soil layers against the periodicity of shortwave radiation emitted by the sun (Luo et al., 2018).

Direct observation of the ground thermal regime can be achieved by taking measurements below the ground surface at various depths, usually in boreholes. There are some benefits to this approach to permafrost monitoring besides the obvious upside of observing permafrost directly. The main benefit is in the effects of the diverse heat transfer processes above the ground surface already being reflected in the below-surface measurements instead of being estimated after the fact based on past observations and soil properties. Spatial variation in ground thermal conditions are attenuated exponentially as depth increases (Williams and Smith, 1990). Logistical considerations are the biggest obstacles to the use of direct observation, as costs and equipment requirements increase with the depth of measurement, making observations at the depths required to effectively monitor permafrost relatively scarce.

1.1.3 Model validation

When employing environmental models, it is important to distinguish between verification and validation. Verification, (derived from the Latin *verus*, meaning true), implies a determination of truth and demonstration of proof, something that is only possible for a closed system that can be fully defined and simulated. Models of open environmental systems cannot achieve this level of certainty (Oreskes et al., 1994). This gap in completeness and certainty when modelling natural processes, for intentional simplicity or for lack of complete understanding of natural mechanisms, introduces error into the results. Estimation errors in model parameters, errors in forcing data or inadequate resolution can lead to systematic distortions of modelled results leading to over- or under-prediction, also known as bias. The use of single point measurements to represent measurements over an area with large variability (Gubler et al., 2011), and the use of land-surface temperature and ground-surface temperature

interchangeably (Luo et al., 2018) are examples of practices in geocryology that may result in the introduction of bias. In their paper on model uncertainty, Wang et al. (2016) examined six state-of-the-art stand-alone land-surface schemes. Their results showed that while good agreement was found with the observation-based 1:4,000,000 Map of the Glaciers, Frozen Ground and Deserts in China's permafrost extent in the Tibetan permafrost area, this did not mean that permafrost dynamics were modelled correctly, and indeed demonstrated various biases in their modelled ground temperatures. They further advocated for detailed model evaluation involving uncertainty quantification analysis of key model parameters to ensure that when models arrive at the right results, they do so for the right reason.

Validation is a demonstration that a model in its domain of applicability yields results within a satisfactory range of accuracy consistent with the intended application of the model (Rykiel, 1996). This definition of validation implies the establishment of performance standards under specified conditions. Validation therefore does not make a statement about the ability of a model to fully and truthfully represent a system, but rather indicates its ability to meet the established performance standards that classify it as fit for a purpose (Rykiel, 1996).

1.2 Goals, aims and objectives

Validation is not an essential activity for evaluating research models but provides an avenue for a modeller to address a model's uncertainties and limitations and to solidly ground models at the science-policy interface (Rykiel, 1996; Eker et al., 2018). Model evaluation at this interface is more about subjective model characteristics not strictly connected to the measured performance of the model, such as costs, simplicity, applicability, intelligibility by the user(s), or how the model is presented (Bennett et al., 2013). Though discussion exists around the relative value of different forms of quantitative and qualitative validation under

specific circumstances, such as low-data situations (Bryant et al., 2018), warnings about the lack of the quantitative type of validation preferred by modellers have been reported for some time. Such warnings include areas of study such as environmental services, climate change, habitat quality of wildlife populations, the distribution of the effect of toxins on wild species, and land change modelling among others (Ochoa and Urbina-Cardona, 2017; van Vliet et al., 2016).

Some efforts to reverse this trend have taken the form of context-agnostic modelling checklists and best practices (e.g. Van Waveren et al., 1999; Jakeman et al., 2006; Council for Regulatory Environmental Modeling, 2009). These tools provide a procedure for the building and evaluation of a model against defined objectives. More thorough exploration of available techniques for characterising the performance of environmental models, such as those by Willmott et al. (2012) and Bennett et al. (2013) delve further into the techniques used in the evaluation of such models, their relative merits, and limitations.

The above tools focus on the validation of a single model with respect to a specific objective. Where there are several models for the same or similar purpose, a decision on the most suitable model for an objective can be difficult for a user. Model developers face similar decisions of whether a new model version represents a tangible improvement over previous versions. Validation methods differ and often involve a means of measuring the accordance between model results and observations using several formulations such as the coefficient of determination (r^2), root mean square error ($RMSE$) and mean bias error (MBE) (e.g. Hinzman et al., 1998; Gent et al., 2011; Badache et al., 2016; Gisas et al., 2016). Such formulations are referred to as accordance measures throughout this work. Thorough assessments on the relative merits of these methods in relation to the model objective is usually excluded from the discussion. This makes gaining insight into the applicability of a model to a user's intended purpose based on comparing validation results difficult. It also makes determination of the relative validity of a model challenging. This difficulty in users and

modellers gaining a full grasp of the modelling landscape for their objectives is a problem that undermines trust in models and hinders the progress of modelling.

The International Land Model Benchmarking (ILAMB) project (Hoffman et al., 2016) is one notable effort to develop a multi-model benchmark that consists of reference data sets, spatial and temporal resolutions, variables and metrics used for the evaluation and comparison of different models and model versions (Collier et al., 2018). ILAMB’s focus is on land models, and not specifically on cold weather dynamics. The ILAMB project has been adapted to benchmark permafrost-related processes through the Permafrost Benchmark System (Piper, 2018). ILAMB provides summary ‘mean state’ analysis, which involves calculation of model mean state in space and time, and its bias and $RMSE$ relative to the observation variable (Collier et al., 2018). While modifications to the system are possible, it is geared towards summary statistics for a narrow set of supported land models against observations and provides no specific tools for further analysis among models.

There are various methods of calculating accordance between observations and modelled results. While r^2 , $RMSE$ and MBE are often used, research on the development of other measures suggest that they are not all equal. Willmott et al. (2012) assert that their refined index of agreement (d_r) is a non-trivial improvement over two previous versions of the index of agreement (d , d_1), as well as over four other comparable dimensionless indices of accordance, while Legates and McCabe (2013) challenge this assertion on behalf of their own modified coefficient of efficiency, E_1 . Both sets of authors do however agree on the notion that any measure of accordance is better served by using seasonal observational data means over those using an entire observational dataset. This raises a question of how much influence the choice of accordance measure has on the determination of a model’s validity for an objective, as well as the effect seasonality and other forms of temporal and spatial granularity have on this determination.

The overarching goal of this research is to improve model development and trust in model

results through validation. Understanding how accordance measures affect determinations of validity can lead to more effective and well-informed use of validity measures matched to specific objectives. Identification of the accordance measure that best evaluates an objective would be valuable in determining how well a model performs against others in a domain, guiding model improvement.

This research aims to investigate the influence that the choice of accordance measure has on the determination of validity of a model for a particular purpose. The performance of multiple models for each observation point in space is quantified using a metric of accordance. Aggregation of these quantifications by model provides an evaluation of the overall performance of each model over the evaluated dataset, and direct comparison of these values yields a ranking for each model. A statistical confidence level of individual model performance can be calculated with model performance over all points. This procedure, when performed with different measures of accordance over temporal and spatial subsets of the data should reveal whether ranks and their confidence levels vary based on the choice of accordance measure and temporal and spatial granularity of the observations data. This aim is accomplished through the following objectives:

The first objective of this thesis is to evaluate the effect of different measures of accordance on determinations of model validity in the prediction of GST by assessing how rankings of the performance of six models over a full year vary based on accordance measure used. The expectation is that rankings will show a small variation through different accordance measures.

The second objective is to evaluate the effect of temporal granularity on rankings of model accordance, and how this varies across accordance measures. At their core, many accordance measures are comparisons of individual model values against a baseline prediction, often the observational mean. Subdividing time-series seasonally localizes their observational mean, providing a better benchmark for comparison (Legates and McCabe,

2013). This localization is especially important in assessing a model's ability to capture the effects of seasonal phenomena such as snow cover and ground thaw on the ground thermal regime. The expectation with this experiment is that two distinct types of rank differences will be found. The first difference is in how each model is ranked across different seasons by the same accordance measure. This difference would highlight the variations in skill of each model at different temperature regimes represented by the seasons. It would also reflect how different error patterns generated by each model/observation comparison are evaluated by each accordance measure. The second is a difference in the model rankings generated by different accordance measures within the same season. This type of difference would indicate that each accordance measure fundamentally evaluates models differently and for the purpose of model selection by ranking, the choice of accordance measure affects which model is ultimately selected.

The third objective extends the second by evaluating the effect of grouping observation points by ground type on seasonal measures of accordance. Ground type grouping allows ranking of model performance over these ground types, and subsequent evaluation of differences in rankings between accordance measures. The expected outcome of this objective is that differences in how a single model is ranked across ground types and season combinations within the same accordance measure will highlight variations in model skill at simulating those specific ground and temperature conditions. The reaction of various ground types to different weather regimes is reflected in the observational dataset, and model accordance values calculated for each ground type by season reflect the skill of each model in simulating these reactions at points. Differences in these model accordance values would serve to highlight the difficulty of extrapolating model validity from one place to another. Differing ranks across accordance measures add further complication to the identification of a best model.

Over all study objectives, the accordances of individual points to model combinations

will be used to assign a level of confidence to a model’s overall mean accordance. This allows for comparison of the confidence in the predictions for each model, and study of how these levels of confidence are affected by the accordance measure, temporal granularity, and spatial grouping selected.

1.3 Methodology

Simulated GSTs are generated using ERA5 and JRA-55 atmospheric reanalysis data interpolated to the location and elevation of each point for which observations are available. Two different configurations of the GEOTop model are combined with the two reanalysis products for four distinct models, and a simple transformation of the 2 m surface air temperature from each reanalysis product yields two additional ad-hoc models so that six different GST simulations are produced per point. Sets of accordance results between simulated GST and observations are calculated for a year, and the overall mean accordance of each model over all points is used to rank the models. This process is repeated with several selected measures of accordance. The use of ranks in this process keeps the focus on relative performance of the candidate models, the true metric of interest for identifying the best model to meet an objective. It is expected that accordance measures will yield different rankings, indicating that the ultimate perceived model skill is partly dependent on the method used to measure that skill. The use of multiple driving data sources and model configurations should enable further exploration of the degree to which rank differences between accordance measures can be linked to either of these inputs.

Modelled and observation time-series are then further subdivided temporally by calculating accordance with observations for each model by season. The resulting model rankings are expected to indicate how model skill varies by season, and how this variation differs across accordance measures. Finally, a spatial aspect is introduced when observation points

are grouped by ground type. Seasonal mean accordances and their ranks are calculated for the models within each ground type. Differences in rankings between the annual, seasonal and ground type groupings may indicate the extent to which model efficacy is affected by the size and homogeneity of the dataset, and thus how model performance on annual or larger datasets may or may not be generalized to subsets within the data.

Confidence intervals are calculated on all annual and seasonal mean accordances with the results allowing direct comparisons of the mean/rank certainty provided by different accordance measures within each step, and how uncertainty varies across temporal and spatial dimensions.

1.4 Scope

Increasing complexity and data requirements of permafrost models are often associated with increased predictive ability. This work is concerned purely with model predictive ability and not model complexity or forcing data requirements. Consequently, interpretations of a model output's accordance with observations will not delve into issues relating to model complexity or equifinality. This complexity also means that the desired output variables normally represent one or a few variables within a simulated environment of many. While the quantification of model error is an important aspect of this work, its scope shall be limited to the evaluation of error within the GST. Error within other simulated environmental outputs will not be evaluated.

The effect of observation error on model selection are beyond the scope of this work. All model outputs tested will be compared against the same observation dataset, and therefore, all accordance measures will evaluate against the same patterns of error regardless of source.

1.5 Structure

Chapter 2 provides background information on important aspects of the research, and its motivation. Chapter 3 outlines the data used as well as the numerical models and their configurations. Chapter 4 presents the accordance measures used for model evaluation, their mathematical properties and the rationale behind their selection. The experimentation to be carried out with these measures to satisfy each objective is presented in Chapter 5. The results of experimentation are detailed in Chapter 6, with interpretation in the context of the research goals presented in Chapter 7. Analysis of the remaining uncertainties of the results, as well as reflection of the results in the context of existing research are presented in Chapter 8. A short conclusion, including potential applications of this research, and future research directions is presented in Chapter 9.

Chapter 2

Background

2.1 Modelling cold environments

2.1.1 Introduction

Simulating geocryological and hydrological phenomena in cold environments presents unique challenges. Water infiltration and freezing in soil from precipitation and snowmelt, and the changing insulating effects of snow accumulation and compaction on the ground are some effects that entail additional model complexity and data requirements on process-based numerical models (e.g. Westermann et al., 2013). These complex dynamics have been integrated into land surface schemes, which when coupled with global or regional atmospheric models are used to forecast weather and to predict climate patterns over large land areas and understand the feedbacks between the ground and atmosphere (e.g. Gent et al., 2011; Nazarenko et al., 2015). Data and computational limitations often mean that the spatial resolution of simulations based on these models is limited to grid squares in the multi-km² range. More detailed models are necessary to study the dynamics of the ground thermal regime where permafrost is present, especially at greater depths and more complex alpine terrains, where

mean annual ground surface temperatures (MAGST) have been found to vary by up to 6°C within a 300 m elevation band, and by up to 2.5°C within a 14 m distance in homogeneous terrain (Gubler et al., 2011). Decoupled land surface schemes, which are run offline and do not affect the driving climate data can be focused on smaller areas or points, and smaller-scale hydrological simulations that integrate cold-weather dynamics can be forced by atmospheric observation or retrospective analysis (reanalysis) products for this purpose.

2.1.2 Driving data

Models are the primary means of studying the evolution of the ground thermal regime in locations underlain by permafrost due to the difficulty and cost of direct observation collection in such remote locations. While the collection of atmospheric data to drive these models is more common, the sparseness of collection points results in large spatial gaps. These gaps can be bridged by downscaling reanalysis products, which provide a multivariate, spatially complete, and coherent record of global atmospheric circulation. Reanalysis products are created by combining available observations with prior information from a forecast model to estimate the evolving state of the global atmosphere and its underlying surface (Dee et al., 2011). The forecast model equations make it possible to extrapolate information from locally observed parameters to unobserved parameters in a physically meaningful way, and to carry this information forward in time (Dee et al., 2011). While this is the general procedure, various reanalysis products exist with differing sets of observations and forecast models, such as the National Centers for Environmental Prediction’s NCEP/NCAR reanalysis (Saha et al., 2010), the Japan Meteorological Agency’s JRA-55 (Kobayashi et al., 2015), The National Aeronautics and Space Administration’s MERRA-2 (Gelaro et al., 2017), and the European Centre for Medium-Range Weather Forecast’s ERA-Interim (Dee et al., 2011) and ERA5 datasets.

As each reanalysis is based on a model, error and uncertainty exist within the data. Though more recent iterations of these models have been demonstrated to be improvements against their predecessors when compared against observations (Reichle et al., 2017; Urraca et al., 2018), general underestimation, and variations in seasonal performance were noted in the soil temperatures of some reanalysis products (Yang and Zhang, 2018). Assessments of reanalysis skill have identified shortcomings in the prediction of clouds and in terrains such as coastal and mountain areas, where spatial grid sizes are inadequate to capture the high variability of some atmospheric variables (Urraca et al., 2018). Gelaro et al. (2017) calculated that the 12-month running mean values of global precipitation in ERA-Interim, MERRA-2, and JRA-55 can differ by almost 20% at times. Reanalysis data presents a valid alternative to observations for models of data-sparse environments, but its limitations should be borne in mind, and more reliable data sources, such as ground and satellite observations used when available (Urraca et al., 2018).

Ground type categorization and accurate data on the hydrological properties of different ground types can significantly improve modelled ground temperatures (Gubler et al., 2013). Reanalysis data describes the conditions above the ground surface. Ground categorization informs composition, conductivity and heat capacity of different ground types, as well as how ground cover affects the rates of energy exchange with the atmosphere. Combined, these data sources provide a means to simulate the ground thermal regime at a point where observations are lacking.

2.1.3 Permafrost modelling

GST is usually a metric of interest for models of the ground thermal regime owing to the ground surface's role as the interface for energy exchange between the atmosphere and the ground. These models can be categorized as process-based models where the thermal state

of the ground is determined based on principles of heat transfer, or empirical and statistical models which can be used to infer permafrost presence, active layer depth, or mean annual ground temperature (Riseborough et al., 2008). As more processes are simulated, model complexity increases, and the data needed to obtain credible results usually increase as well (Riseborough et al., 2008). While reanalysis can be used to fill this data requirement to drive models, calibration and operational validation involves the comparison of model results with data obtained by observation and measurement of the real system (Rykiel, 1996). As such observations are usually limited, a validation procedure must be constrained to an objective that is achievable within the capabilities of the model and the available data.

2.2 Model validity and selection

In the limited technical sense applicable to simulation, validation is an important concept within model evaluation. Conceptual validity refers to the extent to which the model's behavior is consistent with prevailing scientific theory and involves assessments of the correctness of the theories and assumptions underlying the model, along with justifiability of the relationships, structures, simplifications and exclusions within the system (Willmott et al., 1985; Rykiel, 1996). Operational validity refers to the extent to which a model in its domain of applicability yields results within a satisfactory range of accuracy. This range is defined by an objective that reflects the intended application of the model, usually determined by comparisons of model output to observed data (Rykiel, 1996). Operational validation within an evaluation process is often the most tangible means of establishing model credibility, while scientific validation is more intrinsically linked to a specific model or problem, making it harder to generalize (Willmott et al., 1985). Establishing operational validity through a model performance measurement is intrinsically case-dependent, and specific to the field of application, characteristics of the model, data, information and knowledge that we have at

our disposal, and the objectives of the modelling exercise (Bennett et al., 2013). This alignment between model validity and well-defined objectives is important to ensure the correct characteristics are measured in assessing model performance.

2.2.1 Modelling objectives

In the context of model validation, an objective is an expression of the metrics used for assessment, and the thresholds within those metrics that delineate acceptable and unacceptable magnitudes of error between model results and observations. Once a model's results are found to be within the threshold of the objective, the model is qualified for the scenario tested, and can be used to answer the question for which it was constructed (Rykiel, 1996). In defining a validation objective, awareness of what aspects of a system are most important for a model to simulate accurately, and definition of the amount of error tolerable are important. Accordance metrics that quantify model error against observations each emphasize different aspects of the error characteristics. A combination of accordance metrics, including but certainly not limited to *RMSE* and Mean Absolute Error (*MAE*), are often required to assess model performance (Chai and Draxler, 2014).

2.2.2 Current practices in model validation

A literature review of publications related to quantitatively modelling water-related ecosystem services (ES) conducted by Ochoa and Urbina-Cardona (2017) found that 60% (39/65 in-scope publications) did not carry out any validation of the modelling tools they used. This was especially concerning as only four of the nine tools used in these publications were supported by peer-reviewed scientific articles. This lack of validation and peer review mean that the recommendations of those 60% of publications carry a high degree of uncertainty. While there are some, like Rykiel (1996), and more recently van Vliet et al. (2016), who call for a

higher standard of validation in the environmental sciences, and a growing demand for more standardization in evaluation of environmental models (Bennett et al., 2013), objective-based validation remains largely an afterthought in permafrost modelling literature.

Table 2.1 gives a brief overview of publications focused on permafrost modelling and their validation processes. Some of these publications conduct a strictly qualitative validation, consisting of visual comparisons of modelled outputs vs observations (e.g. Osterkamp and Romanovsky, 1999). Simpler quantitative approaches, such as that presented by Guo et al. (2012) conduct direct comparisons of temperatures, in this case change over time of model results to observations. Other forms of quantitative validation are often based on the Root Mean Squared Error (*RMSE*) as a main metric of comparison, but do not discuss the reason for this decision, the impacts of different data patterns on its results, or how the resulting number should be interpreted, a notable oversight given the interpretational challenges of sum-of-squares-based measures (e.g. Badache et al., 2016; Tao et al., 2017). More elaborate validation procedures make use of multiple measures of accordance between model results and observations (e.g. Hinzman et al., 1998; Gissnas et al., 2016). Though Hinzman et al. (1998) specify the purpose of each measure they present, neither example presents methods or thresholds to establish model validity for their purpose. Indeed, among the cited literature, objectives were not included in discussions of validity. Numerical measures of model accordance with observations were included in nearly all studies, but analysis of what these levels of agreement meant, or discussions on what constitutes satisfactory agreement were not. In place of these discussions, vague terminologies such as ‘basically reasonable’ (Guo et al., 2012) and ‘very satisfactory’ (Badache et al., 2016) are often used to describe the results of analysis without indication of what the standard of comparison is.

Table 2.1: Summarized model evaluation procedures of nine permafrost modelling papers. A more detailed table is included as Appendix A.

Model	Evaluation procedure
Community Climate System Model, version 3 (CCSM3) from Lawrence and Slater (2005)	Qualitative map comparison, comparison of Active Layer Thickness (ALT) values
The community climate system model version 4 from Gent et al. (2011)	Qualitative map comparison, comparison of ALT values, comparison of mean of borehole within a gridcell against modelled gridcell values at fixed depths
NASA Catchment Land Surface Model from Tao et al. (2017)	$RMSE$, weighted by layer
CryoGRID 1 from Gisnas et al. (2016)	$RMSE$, r^2 and ME for MAGST and MAGT MAGT: Percentage of modelled points within acceptable deviation from observations reported MAGST: Percentage of observations below 0°C in grid compared to percentage of grid model projections below 0°C
CryoGRID 3 from Westermann et al. (2016)	Qualitative, "satisfactory agreement" cited without numerical definition
The community climate system model version 4 from Guo et al. (2012)	Borehole mean temperature ranges compared against simulated grid means
A new modelling approach for improved ground temperature profile determination from Badache et al. (2016)	Data graphed and qualitatively compared using $RMSE$ and Normalized $RMSE$
Goodrich and the Guymon-Hromadka model from Osterkamp and Romanovsky (1999)	Visual comparison of modeled and observed
Surface energy transfer equation used to drive "subsurface thermal model" from Hinzman et al. (1998)	$RMSE$, MBE , r^2 , Relative Index of Agreement (I_a)

2.3 Accordance

Willmott et al. (2012) described accordance as referring to statistical approaches used to compare model-produced estimates (P) with pair-wise matched reliable observation values (O) in the same unit of measure. These values are often measured and modelled at discrete times and/or locations even though they may in reality be continuous (Willmott et al., 1985), with the j th values of P and O denoted as p_j and o_j respectively, and individual values of their difference ($D = P - O$) expressed as $d_j = p_j - o_j$. These approaches are among the most basic means of assessing model performance in the climactic and environmental sciences and are based on statistical summaries of the model error (Willmott and Matsuura, 2005). The most basic form of these summaries was expressed by Willmott et al. (1985) as:

$$E^{1/y} = \left[\left(\sum_{j=1}^N \omega_j |d_j|^y \right) / \left(\sum_{j=1}^N \omega_j \right) \right]^{1/y} \quad 0 < y \quad (2.1)$$

where y is an exponent used to control error scaling, N is the number of d_j values in D , and ω_j represents a scalar weight applied to $|d_j|$ to compensate for its relative influence within the dataset. ω_j is typically assumed to remain at 1, but can vary in an irregular spatial/temporal observation domain in order to keep $E^{1/y}$ unbiased (Willmott et al., 1985).

2.3.1 Sum-of-squares-based and sum-of-absolute-error-based accordance measures

Two important special cases of Equation 2.1 are:

$$E^{1/y} = MAE \quad y = 1 \quad (2.2a)$$

$$E^{1/y} = RMSE \quad y = 2 \quad (2.2b)$$

MAE and *RMSE* are two commonly employed measures of accordance, however, their use specifically, and the use of sum-of-squares-based measures versus sum-of-absolute values of errors generally is a point of discussion within the model-validation literature. It is contended that sum-of-squares-based error or deviation statistics, far from being unambiguous representations of average error have no clear-cut scientific interpretation of their values. They are, by design, overestimations of average error magnitude whose values are based on the means of individual error magnitudes, the variability among error magnitudes and the number of observations. This sensitivity and over-weighting of extreme error values is generally considered an undesirable characteristic of sum-of-squares-based measures (Willmott et al., 2012). Such measures should therefore be avoided in favor of sum-of-absolute-error or sum-of-absolute-deviation-based measures (Willmott and Matsuura, 2005; Willmott et al., 2009). An argument for sum-of-squares-based metrics contends that while *MAE* is suitable to describe uniformly distributed errors, *RMSE* is more appropriate to represent model performance when the error distribution is expected to be Gaussian and enough samples are present. When evaluating different models using a single accordance metric, differences in the error distributions become more important to differentiating model performance, making a sum-of-squares-based statistic's ability to produce higher variations in values among different sets of model results more desirable (Chai and Draxler, 2014).

2.3.2 Dimensioned vs. dimensionless accordance measures

Equation 2.1 describes the central tendency of error ($D = P - O$) with results unbounded in the positive and negative range in the units of the datasets being evaluated. Another general form for accordance measures can be expressed as:

$$\rho = 1 - \frac{\delta}{\mu} \quad (2.3)$$

where δ is a dimensioned measure of average error magnitude, such as those described by Equation 2.1 and μ is a basis of comparison (Watterson, 1996; Willmott et al., 2012). The selection of δ determines how well error magnitudes are represented and μ determines the sensitivity of ρ to changes in δ , and the lower limit of Equation 2.3. A result of $\rho = 1$ denotes a perfect model when $\delta \geq 0$ and $\mu > 0$. Most instances of accordance measures in this format are defined such that the lower limit of ρ is 0, -1, or $-\infty$ (Willmott et al., 2012). In absolute-error-based statistics, the values of δ and μ are often set to varying forms of the *MAE*, a measure of the distance between individual observations and their corresponding predictions, and Mean Absolute Difference (MAD), a measure of variation between individual observation points about their mean (Tables 2.2, 2.3). The resulting dimensionless quantity is not easily interpreted in a physically meaningful way and meaning differs by statistic even for those constrained to the same ranges. This is due to the varying responses of various statistics to similar data patterns, a topic that has been widely investigated in literature (e.g. Willmott et al., 2012; Bennett et al., 2013; Legates and McCabe, 2013; Willmott et al., 2015). The advantage of such measures is mainly that their assessments of model performance are scaled to fixed ranges, and therefore interpretation of these assessments and comparisons of models do not require in-depth knowledge of the underlying data and its

Table 2.2: Dimensioned measures of accordance. From Bennett et al. (2013) and Willmott et al. (2012).

Measure	Formula	Basis
Mean Squared Error (MSE)	$\frac{\sum_{j=1}^n (p_j - o_j)^2}{n}$	Sum-of-Squares
Root Mean Squared Error (RMSE)	$\sqrt{\frac{\sum_{j=1}^n (p_j - o_j)^2}{n}}$	Sum-of-Squares
Mean Absolute Error (MAE)	$\frac{\sum_{j=1}^n p_j - o_j }{n}$	Sum-of-Absolutes
Mean Absolute Deviation (MAD)	$\frac{\sum_{j=1}^n o_j - \bar{o} }{n}$	Sum-of-Absolutes
Mean Bias Error (MBE)	$\frac{\sum_{j=1}^n (p_j - o_j)}{n}$	Other

amplitudes.

The differing strengths of error statistics lead much of the literature on the topic to conclude that an adequate understanding of the error in the data is only possible through careful selection of a measure that is appropriate for what is being studied, or through the use of various measures that, when considered together, identify the error most relevant to the chosen objective (Willmott et al., 1985; Legates and McCabe, 1999; Bennett et al., 2013; Legates and McCabe, 2013).

Table 2.3: Dimensionless measures of accordance. For all measures listed, a result of 1 denotes a perfect model. $\bar{o}$ is weighted mean direction or vector of elements in $O \left(\frac{\sum_{j=0}^n \omega_j o_j}{\sum_{j=0}^n \omega_j} \right)$, ω_j is a scalar weight to compensate for an element's relative influence within the dataset, and is typically assumed to remain at 1. c is a scaling constant, preferentially set to $c = 2$ for d_r that controls the influence of the Mean Absolute Deviation portion of the fraction. r^2 from Fox et al. (1961), d from Willmott (1981), d_1 and d_r from Willmott et al. (2012), $E(NSE)$ from Nash and Sutcliffe (1970), E_1 from Legates and McCabe (1999).

Measure	Formula	Range
r^2	$\frac{\left(\sum_{j=1}^n (o_j - \bar{o})(p_j - \bar{p}) \right)^2}{\sum_{j=1}^n (o_j - \bar{o})^2 \cdot \sum_{j=1}^n (p_j - \bar{p})^2}$	0 to 1
d	$1 - \frac{\sum_{j=1}^n \omega_j p_j - o_j ^2}{\sum_{j=1}^n \omega_j (p_j - \bar{o} + o_j - \bar{o})^2}$	0 to 1
d_1	$1 - \frac{\sum_{j=1}^n \omega_j p_j - o_j }{\sum_{j=1}^n \omega_j (p_j - \bar{o} + o_j - \bar{o})}$	0 to 1
d_r	$\begin{cases} 1 - \frac{\sum_{j=1}^n o_j - p_j }{c \cdot \sum_{j=1}^n o_j - \bar{o} } \text{ when } \sum_{j=1}^n p_j - o_j \leq c \cdot \sum_{j=1}^n o_j - \bar{o} \\ \frac{c \cdot \sum_{j=1}^n o_j - \bar{o} }{\sum_{j=1}^n o_j - p_j } - 1 \text{ when } \sum_{j=1}^n p_j - o_j > c \cdot \sum_{j=1}^n o_j - \bar{o} \end{cases}$	-1 to 1
$E(NSE)$	$1 - \frac{\sum_{j=1}^n (o_j - p_j)^2}{\sum_{j=1}^n (o_j - \bar{o})^2}$	$-\infty$ to 1
E_1	$1 - \frac{\sum_{j=1}^n o_j - p_j }{\sum_{j=1}^n o_j - \bar{o} }$	$-\infty$ to 1

2.3.3 Purpose and ranges of accordance measures

Measures of accordance reduce comparisons between simulation results and observations to a single number that represents the deviation detected between the two, and the range to which this resulting number is mapped varies by measure. A ranking framework between models mitigates an immediate need to fully understand the range and character of each accordance measure beyond whether a lower or higher value is more desirable. Even so, it is important to understand the range of each measure in order to define performance objectives for models and for users to decide whether differences in model performance are reflective of the cost associated with running them. Some commonly-used accordance measures are detailed in the following subsections.

2.3.4 RMSE and MAE

The *RMSE* and the *MAE* are sum-of-squared error-based and sum-of-absolute error-based accordance measures respectively, and both are dimensioned, ranging in value from 0 to ∞ . Their equations are listed in Table 2.2 and their relationship is detailed in Section 2.3.1. While *RMSE* is one of the most widely reported error measures in climatic and environmental literature, it has no consistent functional relationship between RMSE and average error (Willmott and Matsuura, 2005), and is therefore not simple to interpret as is often assumed. The non-absolute version of *MAE*, known as *MBE* is sometimes reported to convey average model bias, but should be interpreted carefully as it is inconsistently related to typical-error magnitude, other than being an underestimate due to cancellation between negative and positive error values (Willmott and Matsuura, 2005). For the purpose of ranking model performance without regard to direction of bias, *MBE* does not convey any information not contained in *MAE*, and it therefore not included in this work.

2.3.5 Coefficient of determination (r^2)

The Coefficient of Determination, or r^2 , is a dimensionless sum-of-squares-based accordance measure with a range of 0 to 1, where 1 denotes a perfect model. It is based on the Pearson Product Moment Correlation, also known as r , which is sample covariance between prediction and observations divided by the product of their respective standard deviations. Formulas for both are in Table 2.3. r measures co-linearity between the observations and predictions being compared, while its square, r^2 describes the proportion of the total variance explained by the model (Willmott, 1981). While r^2 was at one time the preferred measure of model performance used by geographers, its role as a sole accordance measure has declined due to its shortcomings in fully describing model error. The main shortcoming is its standardization for differences between observed and predicted means, meaning that offsets between values are not accounted for, sometimes resulting in high scores for models that predict patterns, but do not necessarily predict the amplitude or correct average value of an environmental variable. r^2 results will therefore not be further considered in this work.

2.3.6 Nash-Sutcliffe efficiency (NSE)

A dimensionless, sum-of-squares-based measure, NSE was initially devised by Nash and Sutcliffe (1970) to aid in parameter optimization of hydrological models simulating river flows resulting from rainfall. NSE can be conceptually defined as the ratio of the mean square error to the variance of the observed data, subtracted from unity (Legates and McCabe, 1999). More simply put, the NSE compares the performance of a model to a benchmark model that only uses the mean of the observed data. Its range is $-\infty$ to 1, where 1 represents a perfect model, 0 indicates performance no better than using the observed mean as a model (Bennett et al., 2013), and negative values indicate that the model has failed to explain more of the variability in the observations than their mean (Legates and McCabe,

2013). Values between 0 and 1 are interpreted as the percent of variance in the observations that is explained by the model predicted values (Legates and McCabe, 2013). The formula for NSE can be found on Table 2.3. *NSE* presents an improvement over r^2 in that its value reflects differences in both mean and variance, but still has some drawbacks. Apart from the shortcomings associated with sum-of-squares-based measures, *NSE* has a tendency towards assigning higher values for timeseries that display higher dynamics such as fluctuations between high temperatures, and lower values for lower dynamics such as temperatures hovering about 0. In order to obtain comparable NSE values to a high-dynamics timeseries, a low-dynamics timeseries must have a better prediction (Krause et al., 2005).

2.3.7 Modified coefficient of efficiency (E_1)

Formulated by Legates and McCabe (1999), E_1 is a dimensionless sum-of-absolutes-based accordance measure. A generic form of NSE can be written as:

$$E_j = 1 - \frac{\sum_{j=1}^n |p_j - o_j|^j}{\sum_{j=1}^n |o_j - \bar{o}|^j} \quad (2.4)$$

where $j = 2$, making NSE equal to E_2 . The reformulation of Equation 2.4 with $j = 1$ gives E_1 (see Table 2.3), which can more simply be expressed as the ratio between the Mean Absolute Error of the model and the Mean Absolute Deviation within the observations subtracted from one. Legates and McCabe (2013) argue that this is preferable to the NSE because of the decreased weight placed on outliers. Like NSE, the range of values is from $-\infty$ to 1, where 1 represents a perfect model and 0 indicates performance no better than using the observed mean as a model. Values between 0 and 1 represent the percentage of the absolute value difference a model can explain between the observations and model predictions.

2.3.8 Willmott's index of agreement (d)

First introduced by Willmott (1981), the first version of Willmott's index of agreement (d) was meant to be easily interpreted and to enable cross-comparisons of its magnitudes for a variety of models, regardless of units (Willmott, 1981). It is a dimensionless, sum-of-squares-based accordance measure that describes the relative covariability of predictions and observations about an estimate of the 'true' mean, in this case the observational mean (Willmott et al., 2012). d operates within a range of 0 to 1, where 1 indicates perfect agreement and 0 connotes one of a variety of complete disagreements (Willmott, 1981). At the time of its introduction, Willmott contended and demonstrated that due to its sensitivity to differences between the observed and predicted means as well as to certain changes in proportionality, d represented an improvement on r and r^2 . Though it is an improvement, d still has its own shortcomings, namely, its tendency towards relatively high values (> 0.65), even for poor model fits, leaves only a narrow range for model calibration, and despite this, it is not sensitive to systematic model over- or under- prediction (Krause et al., 2005).

2.3.9 Willmott's modified index of agreement (d_1)

Willmott et al. (1985) introduced d_1 as an update on d . Recognizing the problems inherent with sum-of-squares-based measures, d_1 is a dimensionless sum-of-absolutes-based measure. Its formula, is based on a generalized expression of d as:

$$d_y = 1 - \frac{\sum_{j=1}^n \omega_j |p_j - o_j|^y}{\sum_{j=1}^n \omega_j (|p_j - \bar{o}| + |o_j - \bar{o}|)^y} \quad (2.5)$$

where ω_j is a scalar weight to compensate for an element's relative influence within the dataset that is typically assumed to remain at 1 (Willmott et al., 1985). Under this equation, the special case when $y = 2$ (d_2) is the original form of d , and $y = 1$ is the updated d_1 . Its

range and interpretation, like d , are from 0 to 1, with 1 indicating perfect agreement and 0 connoting one of a variety of complete disagreements. Noted advantages of d_1 over d are that it approaches 1.0 more slowly as predictions approach observation values, providing greater separation when comparing models that perform relatively well, and it is also much less sensitive to the shape of the error-frequency distribution and, as a consequence, to errors concentrated in outliers (Willmott et al., 2012).

2.3.10 Willmott's refined index of agreement (d_r)

The most recent form of Willmott's Index of Agreement, d_r is a dimensionless, sum-of-absolutes-based accordance measure. This version of the index is touted as a nontrivial improvement over earlier iterations that is quite flexible and therefore applicable to an extremely wide range of model-performance applications (Willmott et al., 2012). The two major differences between d_1 and d_r are the removal of the influence of model predictions on the denominator (μ) of the d_1 formula, and the extension of the measure's lower range from 0 to -1. The removal of predictions from the denominator reflects the view that as long as the denominator contains model results, it cannot be a true model-independent basis of comparison for the numerator (δ). This change was achieved through the replacement of prediction values in d_1 's denominator with observation values, with the revised d_1 (d'_1) written as

$$d'_1 = 1 - \frac{\sum_{j=1}^n |p_j - o_j|}{2 \cdot \sum_{j=1}^n |o_j - \bar{o}|} \quad (2.6)$$

This revision has the effect of unbounding the lower end of d'_1 , undermining the interpretation of values associated with poorly performing models (Willmott et al., 2012). For this reason, the lower range is redefined by inverting the fractional part of the expression and subtracting 1.0 from it, when d'_1 falls below zero, for the final form of d_r as presented in Table

2.3. A d_r score of 1 indicates a perfect model, while numbers in the positive range indicate a comparison between sum of the magnitudes of the differences between model predictions and observations, and the sum of magnitudes of deviations of a theoretical perfect model (predictions = observations) about the observational mean. A value of 0.5 represents equivalency between the two differences, indicating that the similarity of the predicting model to the perfect model is no better than the observational mean. Negative values of d_r indicate sum of magnitude of deviation between observation and model values greater than those of the perfect model about the mean, with $d_r = -0.5$ indicating a ratio of 2:1 between these values, though interpretation of negative values as d_r moves closer to -1 do not have such clear-cut interpretations and should be done with caution (Willmott et al., 2012). A specific criticism of d_r 's interpretability revolves around the special meanings of $d_r = 0.5$ in contrast to E_1 , a similar measure that has its equivalent value at 0, and the resulting high values assigned to even relatively poor models, a criticism that was also made of the original index of agreement (Legates and McCabe, 2013). Additionally, the added complexity of remapping the negative part of the results range is criticized as unnecessary since d_r 's values below 0.5 provide only description of a model's 'level of inefficacy', a portion of the statistic that is largely considered immaterial (Legates and McCabe, 2013).

Chapter 3

Data and Models

3.1 Introduction

This thesis experimentally studies the effect that different methods of calculating accordance between GST observation and model results have on the competence attributed to each model. This chapter describes the observations data, the reanalysis data that drives the models, and the models themselves.

3.2 Observations

Observations used in this work are ground temperature measurements collected by Gruber et al. (2018). This dataset consists of observations collected in the vicinity of Lac de Gras, NWT, Canada. Measurements were taken within 15 m x 15 m plots selected to represent different facets of the landscape, and within each plot, up to four subplots to represent the diverse surface characteristics within the plot area (Gruber et al., 2018). GST was recorded at each subplot using single-channel GeoPrecision M-Log5W-SIMPLE temperature loggers, which operate at a resolution of 0.01°C with a long-term stability of $\pm 0.1^\circ\text{C}$, and an accu-

racy of $\pm 0.1^{\circ}\text{C}$ at 0°C (Gruber et al., 2018). The sensors were placed at a depth of 10 cm and set to record at 20-minute intervals between August 2015 and September 2017. In all, timeseries data on ground surface temperature for 159 observation points are used, each with varying amounts of data between 2015 and 2017, and summarized as daily means. Where gaps greater than one day are present within a dataset, the containing time-period (year or season) is discarded. The ground type classification process used by Cao et al. (2019) groups each point into one of 10 ground types that are defined based on surface and subsurface characteristics, such as surficial geology, vegetation type and height and topography. These characteristics were chosen because they explain a significant part of the observed spatial variation of GST, and are easy to differentiate in the field, though they introduce a degree of subjectivity into the classifications. The ground types and their main characteristics are described in more detail in Appendix B.

In all, 489 annual timeseries were obtained from the dataset and of these, 213 were discarded due to missing days within the timeseries, leaving 276 complete year timeseries. Figure 3.1 illustrates the distribution of the number of days present within the incomplete years.

When generating the seasonal datasets, the incomplete seasons usually found at the start and end of each point's timeseries were removed. An additional 94 seasonal timeseries were found to be missing at least one day and discarded. A breakdown of the resulting dataset of complete seasons is included as Table 3.1, and the numbers when seasons are grouped by ground type are presented in Table 3.2.

Figure 3.1: Graph of the distribution of the number of days present in the 213 incomplete years in the dataset.

Table 3.1: Seasonal start dates and number of complete seasons of point observations data available.

Season	Start date	Count
Summer	01-Jun	282
Fall	01-Sep	252
Winter	01-Dec	297
Spring	01-Mar	294
Total		1,125

Table 3.2: Number of observation points and their ground types.

Ground type	Number of observation points	Number of seasonal timeseries
Esker	14	88
Ice wedge polygon	13	99
Ice wedge trough	11	70
Sedge fen	4	32
Sedges and grass	28	203
Short vegetation	15	110
Shrubby hummocks	27	191
Snowdrift	12	91
Tall shrubs	4	32
Till veneer	31	209
Total	159	1,125

3.3 Reanalysis data

3.3.1 GlobSim

GlobSim is an open-source software toolkit to facilitate working with ensembles of reanalysis data from various sources by standardizing data structures and units, interpolating and scaling data covering an area to discrete points, and outputting the result to a common format (Cao et al., 2019). Using GlobSim, ERA5 and JRA-55 reanalysis data were downloaded, interpolated and scaled to specific points at an hourly resolution.

3.3.2 ERA5

Created and maintained by the European Centre for Medium-Range Weather Forecasts (ECWMF), ERA5 is the fifth generation of the ECMWF atmospheric reanalyses of the global climate (Hersbach and Dick, 2016). It is the successor and eventual replacement to the popular ERA-Interim reanalysis. Improvements and innovations over previous iterations include higher spatial and temporal resolutions and uncertainty estimates for the data (Hersbach and Dick, 2016). ERA5 covers the period from 1979-present, and is updated in near real-time with an hourly resolution for analysis and forecast fields. Its spatial resolution is 31 km globally (Hersbach et al., 2019).

3.3.3 Japanese 55-year reanalysis (JRA-55)

Completed by the Japanese Meteorological Agency (JMA) in March 2013, JRA-55 is the second Japanese global atmospheric reanalysis, following the Japanese 25-year Reanalysis (JRA-25). Its main objectives were to address issues found in previous reanalyses and to produce a comprehensive atmospheric dataset suitable for studying multidecadal variability

and climate change (Kobayashi et al., 2015). JRA-55 features considerable progress in mitigating biases identified in JRA-25 through an improved data assimilation scheme and the use of more data sources, though work is ongoing to further mitigate these and other biases still present. It covers the period starting from 1958 and since its release has operated on a near-real-time basis, outputting meteorological variables with a 6-hourly temporal resolution and a ~ 55 km horizontal resolution (Kobayashi et al., 2015).

3.3.4 Use in model validation

ERA5 and JRA-55 reanalysis data were interpolated to the points being simulated and used as meteorological forcing to drive GST simulations in the GEOTop model. ERA5 and JRA-55 were selected for this study due to their spatial completeness and temporal coverage of the domain under study, and the ease of access provided by the GlobSim software. Though GlobSim also supports ERA-Interim and MERRA-2 datasets, assessments of monthly climatological data conducted by Simmons et al. (2017) conclude that ERA-Interim and JRA-55 both perform more consistently with each other and with the conventional surface temperature datasets than MERRA-2, and that MERRA-2 substantially underestimates the recent warming seen in all the other datasets they considered. Furthermore, ERA5 has been found to achieve a reduction of 50%-75% in average bias for surface irradiance over both ERA-Interim and MERRA-2 (Urraca et al., 2018). The variables required by GEOTop as forcing, and derived through reanalysis are listed in Table 3.3.

Table 3.3: Meteorological variables required by GEOtop, and their units/formats.

Dimension	Unit/Format
Date/Time (Date)	DD/MM/YYYY HH:MM
Precipitation Intensity (IPrec)	mm h ⁻¹
Wind Velocity (WindVelocity)	m s ⁻¹
Wind Direction (WindDirection)	°N
Relative Humidity (RH)	%
Air Temperature at 2 m above ground surface (AirTemp)	°C
Air Pressure (AirPress)	bar
Short Wave Solar Global Radiation (SWglobal)	W m ⁻²
Long Wave Incoming Radiation (LWin)	W m ⁻²

3.4 Models

3.4.1 The GEOtop model

GEOtop 2.0 is a physically based simulator of the evolution in time of temperature and water content in the soil, and snow cover. It uses meteorological forcing data to drive mathematical descriptions of terrain effects (topography, vegetation), energy and water balance, and uses atmospheric forcing data to establish boundary conditions that describe interactions with the surface energy and water fluxes (Endrizzi et al., 2014). GEOtop 2.0 employs a volumetric system to simulate soil layers of differing thicknesses. This system resolves the heat equation while accounting for soil thermal conductivity, liquid water and ice content, the soil freezing characteristic curve based on soil water retention curve, and soil water flow. Above-ground, GEOtop simulates the heat flux exchanged with the atmosphere on simple and complex terrain. This is calculated using net shortwave (solar) radiation, net longwave radiation, and turbulent fluxes of sensible and latent heat (Endrizzi et al., 2014). GEOtop's simulation of snow cover dynamics include snow warming and cooling, melting and refreezing, water

percolation, accumulation due to snow precipitation, deposition of wind-blown snow, erosion due to wind and densification due to snow metamorphism (Endrizzi et al., 2014). GEOTop's coverage of the spectrum of cryospheric processes are of particular interest for this work.

3.4.2 GEOTop model configuration parameters

GEOTop is a modular system that allows for variations in model structure based on which modules are enabled. The resulting changes in parametrization of aspects of the energy balance in complex terrain, and snow and vegetation effects are used to create different models within the GEOTop simulation.

The blowing snow module allows GEOTop to account for the impact of blowing snow on vegetation. GEOTop's implementation of blowing snow is based on the Prairie Blowing Snow Model (PBSM) by Pomeroy et al. (1993). PBSM simulates snow transport via the movement of particles in a 'skipping' action just above the snow surface, known as saltation, and the movement of particles suspended by turbulence in the atmospheric layer extending to the top of the surface boundary-layer, known as suspension Pomeroy et al. (1993). Snow sublimation rates calculated from saltating and suspended blowing snow are used in calculations of snow surface erosion and accumulation within GEOTop. A simplified parameterization of snow densification due to wind drifting based on PBSM is implemented for one-dimensional simulations in GEOTop. This complements the regular densification based on Anderson (1976), Jordan (1991) and Jordan et al. (1999). The blowing snow module is used to create two GEOTop models, one employing its effects and one not. The comparison of the results of these two models is not meant to evaluate the efficacy of the different model structures, but rather to present an analysis of how the relative efficacy of the different models is evaluated by accordance measures.

3.4.3 Parameters representing ground types

Tao et al. (2017) demonstrated that enhanced treatments of meteorological forcing, land cover type, and organic carbon-related soil thermal properties can substantially reduce model error, so model parameters and initial conditions to represent the ten different ground types identified were based on field observations, as noted in Section 3.2.

3.4.4 Simple models

In addition to the GEOtop models, a simple model of ground surface temperature, defined as half the surface air temperature daily average for the driving reanalysis data is also used. Its purpose is to help analyze how the relative efficacy of different models is evaluated by different accordance measures.

3.4.5 Resulting models

The components of this section were combined to generate the six models used in this work. Four models were generated by using ERA5 and JRA-55 reanalysis data to drive GEOtop configurations with and without the blowing snow module turned on. A further two models were the simple models described in 3.4.4 based on the 2 m temperatures provided by ERA5 and JRA-55. Table 3.4 provides abbreviated names used to reference these models throughout the rest of this work.

Table 3.4: *Model names, referencing numbers, and description for models used in experimentation.*

Number	Abbreviation	Model Description
1	GT-ERA5	GEOTop, blowing snow module OFF + ERA5
2	GT-JRA55	GEOTop, blowing snow module OFF + JRA-55
3	GT-BSnow-ERA5	GEOTop blowing snow module ON + ERA5
4	GT-BSnow-JRA55	GEOTop blowing snow module ON + JRA-55
5	ERA5/2	ERA5 Air temperature / 2
6	JRA55/2	JRA-55 Air temperature / 2

Chapter 4

Methods

4.1 Introduction

This chapter describes the data processing and modelling steps taken to enable experimentation, the ranking framework used to evaluate models, and the selection process for the accordance measures used to evaluate models in this thesis.

4.2 Model runs

The first step towards experimentation was generating simulation data to compare observations against. Coordinates from the observations data were provided in decimal degrees at a precision of 6 decimal places (resolution of about 0.04 m at Lac de Gras). The four GEOTop models described in Section 3.4 were run using reanalysis data interpolated to the observation locations and pressure levels scaled to the observation elevations as atmospheric forcing. The model soil, snow and vegetation parameters presented in Appendix B were matched to each simulated point by the ground type of the corresponding observation point. Each observation point therefore had four GEOTop configuration files reflecting common soil, snow

and vegetation parameters, but different combinations of reanalysis source and status of the blowing snow module. For each of the four model simulations per point, the GEOtop model was run seven times using the reanalysis data from between January 2000 and January 2015 as spin up, and then run once from January 2015 to August 2018 to obtain the simulation results. Observations and simulation results were then temporally aggregated to daily means. The two simple models entailed dividing each of the two reanalysis 2 m air temperature timeseries at the interpolated observation locations by two and temporally aggregating the result to daily means. The resulting daily observation and six simulation time-series for all 159 points comprise the complete dataset used.

4.3 Model ranking

This thesis uses accordance measures to reduce a comparison of a GST observation and model simulation time-series to a single number, facilitating comparisons of the performance of various models over the same time-period. The mean accordance of a model over all individual points, denoted as $\bar{x}$, is used to summarize that model's performance over the entire dataset. Selection of the best-performing model from a group of models over the observations dataset can be achieved by simply ranking each model's $\bar{x}$ against that of other models in the group, with ranking direction determined by the interpretation of each accordance statistic. This simple ranking framework allows the study of variations in rankings produced by different accordance measures used on the same models and datasets.

4.4 Accordance measure selection

The relative validity of models within each accordance measure is the main metric of interest in this work. Though the ranges and values output by each accordance measure usually differ,

the rankings of models across measures often follow similar patterns.

In this section, measures that present similar rank patterns are grouped, and some causes for these similarities in each group are deduced based on relationships between the formulations of the accordance measures concerned. An accordance measure that is representative of each group is then chosen for experimentation. This reduced number of accordance measures allows for analysis that focuses on measures with strongly differing model rank results, which are the focus of this study. The groupings and representative accordance measures are shown on Table 4.1.

Table 4.1: *Accordance measure groupings based on model rankings and common basis. Measures are grouped vertically. Bold text denotes the measure selected for experimentation from each group.*

RMSE	d	NSE
MAE	$\mathbf{d_1}$	d_r
		E_1

Although strictly a result and therefore more suitable for Section 6, Table 4.2 is shown here to explain the narrowing and simplification of experimentation to be carried out. Variations in the rankings within accordance measures indicating a difference in perceived model performance on annual observations data can be seen in this table.

Table 4.2: *Model annual accordance ranks and means as calculated by different accordance measures.*

Model	MAE	$RMSE$	NSE	d	d_1	d_r	E_1
GT-ERA5	2 / 2.2039	1 / 2.8081	2 / 0.7615	1 / 0.9542	1 / 0.8173	2 / 0.7964	2 / 0.5927
GT-JRA55	3 / 2.2731	3 / 2.9088	3 / 0.7596	3 / 0.9483	3 / 0.8039	3 / 0.796	3 / 0.5919
GT-BSnow-ERA5	4 / 3.1848	4 / 4.0699	4 / 0.4537	4 / 0.9227	4 / 0.7683	4 / 0.6955	4 / 0.3909
GT-BSnow-JRA55	1 / 2.1968	2 / 2.8237	1 / 0.7664	2 / 0.9524	2 / 0.8138	1 / 0.8006	1 / 0.6013
ERA5/2	6 / 4.2609	6 / 5.4148	6 / -0.0263	6 / 0.8452	6 / 0.6718	6 / 0.5863	6 / 0.1695
JRA55/2	5 / 4.1902	5 / 5.2461	5 / 0.0858	5 / 0.8516	5 / 0.6723	5 / 0.5976	5 / 0.1931

While ranks remain stable for the four models at the bottom four rankings (GT-JRA55, GT-BSnow-ERA5, JRA55/2, ERA5/2 ranked 3 to 6 respectively), the top two models display notable variation. The two annual rank orders of GT-ERA5 and GT-BSnow-JRA55 were 1, 2 by $RMSE$, d and d_1 , and 2,1 by MAE , NSE , d_r and E_1 . These groupings are significant as the literature on accordance measures focuses on the difference between sum-of-squares-based and sum-of-absolutes-based measures, with the consensus being that sum-of-squares-based measures have their place, but sum-of-absolutes-based measures are preferable for model evaluation (Legates and McCabe, 1999; Willmott and Matsuura, 2005; Willmott et al., 2009). The expected division between accordance measures in this study was therefore between sum-of-squares and sum-of-absolutes-based, but this is not the case shown by Table 4.2. The connection between d and d_1 , and NSE , d_r and E_1 is instead their common basis in d and NSE respectively. The formulations for NSE , d_r and E_1 share a common denominator in their positive ranges that is the squared or absolute form of $\sum_{j=1}^n o_j - \bar{o}$, which is intended to eliminate the influence of predictions on the number those very same predictions are compared against. d_1 is simply a sum-of-absolutes-based reformulation of the sum-of-squares-based d , and these measures yield the same rankings. This is the same relationship shared by NSE and E_1 , while d_r 's range from 0.0 to 1.0 is defined mathematically as E_1 rescaled by $\frac{1}{c}$, where 2 is the recommended value of c . This relationship is easily observable in the histogram for these two measures at Figure C.7 and Figure C.8.

The rankings for $RMSE$ and MAE also differed for these two top-ranked models, with their difference in error being no greater than 0.02°C , but the difference in value between the top-ranked models of both measures being a larger 0.6°C , indicating $RMSE$'s tendency to exaggerate error.

The ranks of all models were also calculated over four seasons and the results are found in Tables 4.3 through 4.6. The ranks assigned to each model within a season by $RMSE$

and MAE were identical. More generally, the ranks within each season remained mostly constant for all measures in the fall and winter (best and worst rated seasons by mean of all models respectively), with slight variations between the ranks of only the top two winter and bottom two fall models. These variations are in line with previous findings of d and d_1 ranks differing from NSE , d_r and E_1 ranks.

Table 4.3: Models ranked by mean seasonal accordance over all points for summer.

Model	MAE	$RMSE$	NSE	d	d_1	d_r	E_1
GT-ERA5	4	4	4	2	4	4	4
GT-JRA55	6	6	6	4	3	6	6
GT-BSnow-ERA5	3	3	3	1	1	3	3
GT-BSnow-JRA55	5	5	5	3	2	5	5
ERA5/2	2	2	1	6	6	2	2
JRA55/2	1	1	2	5	5	1	1

Table 4.4: Models ranked by mean seasonal accordance over all points for fall.

Model	MAE	$RMSE$	NSE	d	d_1	d_r	E_1
GT-ERA5	1	1	1	1	1	1	1
GT-JRA55	2	2	2	2	2	2	2
GT-BSnow-ERA5	4	4	4	4	4	4	4
GT-BSnow-JRA55	3	3	3	3	3	3	3
ERA5/2	5	6	6	6	5	5	5
JRA55/2	6	5	5	5	6	6	6

Within their group, d and d_1 showed some variation in the summer, with a smaller variation in fall and spring, and no variation in winter. As the differences in rank between d and d_1 show no obvious pattern, a cause is difficult to ascertain, though the difference in error weighing between sum-of-squares and sum-of-absolute-based statistics would be the most

Table 4.5: Models ranked by mean seasonal accordance over all points for winter.

Model	<i>MAE</i>	<i>RMSE</i>	<i>NSE</i>	<i>d</i>	<i>d</i> ₁	<i>d</i> _{<i>r</i>}	<i>E</i> ₁
GT-ERA5	3	3	3	3	3	3	3
GT-JRA55	1	1	1	2	2	1	1
GT-BSnow-ERA5	4	4	4	4	4	4	4
GT-BSnow-JRA55	2	2	2	1	1	2	2
ERA5/2	6	6	6	6	6	6	6
JRA55/2	5	5	5	5	5	5	5

Table 4.6: Models ranked by mean seasonal accordance over all points for spring.

Model	<i>MAE</i>	<i>RMSE</i>	<i>NSE</i>	<i>d</i>	<i>d</i> ₁	<i>d</i> _{<i>r</i>}	<i>E</i> ₁
GT-ERA5	2	2	3	1	1	2	3
GT-JRA55	3	3	1	3	4	3	2
GT-BSnow-ERA5	4	4	4	4	3	4	4
GT-BSnow-JRA55	1	1	2	2	2	1	1
ERA5/2	5	5	6	5	5	5	6
JRA55/2	6	6	5	6	6	6	5

obvious hypothesis.

Within the *NSE*, *d*_{*r*} and *E*₁ group, variations in summer and fall are restricted to the ERA5/2 and JRA55/2 models, with sum-of-absolutes-based measures (*d*_{*r*} and *E*₁) agreeing on ranks, but disagreeing with the sum-of-squares-based measure (*NSE*). There is no variation in winter ranks within this group. Spring displays the most variation, including a difference in rank between *d*_{*r*} and *E*₁, a rarity due to their common formulation within the positive range. This variation could be due to the nature of the spring season, which contains both negative and positive temperature periods, the persistence of a nearly constant temperature very close to the freezing point during annual thawing of the active layer known as

the zero curtain (Van Everdingen, 2005), and the differences in how these measures scale, resulting in negative values. Further study is needed to establish the cause conclusively.

Though the grouping patterns can be inconsistent when accordance values delve into the negative end of the accordance measure ranges, ranking frameworks are more often concerned with identifying best models, so this inconsistency among poorly performing models is not as notable.

Overall, these results indicate that a selection of one accordance measure per group identified will be enough to illustrate rank variations present within this dataset. *RMSE* will be used over *MAE* due to its ubiquity within the literature. Willmott's d_1 is presented as an improvement on d (Willmott et al., 1985), and will therefore be used in its favour. Among *NSE*, d_r and E_1 , *NSE* is a poor choice for environments that experience freezing and thawing cycles due to its amplification of error, tendency to assign lower values for lower dynamics, and unbounded negative range. This can be seen in the mean seasonal *NSE* values over all observation points in the Gruber et al. (2018) dataset are shown in Table D.1 in the appendix. At -244.3 , the *NSE* value for winter is two orders of magnitude larger than the next lowest negative value given by *NSE* (-6.8) in the table. Locations or ground types that remain close to 0°C in Arctic or sub-Arctic environments, such as those that are well insulated by snow, or have high water content and experience long periods of freeze or thaw will experience extended periods of low dynamics. Model error within low dynamics is amplified far more than similar error over medium or high dynamics, resulting in extreme negative values. An example of a low dynamic data pattern that results in extreme values is included as Figure 4.1. Contrast this against Figure 4.2, which is a similar data pattern that also results in large negative *NSE* values, but remains two order of magnitude lower, demonstrating the effect proximity to 0°C has on the negative growth of *NSE*. *NSE* should therefore be avoided in permafrost research and temperature validation scenarios where low dynamics are likely to be encountered.

Figure 4.1: Graph of ground surface temperature on tall shrubs at point *Bushloggers_BS03* in winter 2015, demonstrating a temperature pattern that results in extreme negative *NSE* values. The winter season is highlighted in green. In this case, the lowest *NSE* value is -127,125.28 for the *ERA5/2* model, and the highest is -4257.4 for the *GT-JRA55* model.

Though d_r and E_1 are based on *NSE*, both mitigate its issues by being sum-of-absolutes-based, and are therefore quite similar in their positive ranges. This is seen in the E_1 values in Table D.1, where a similar pattern to *NSE* of a winter value proportionally much lower than that of other seasons is observed, though the magnitudes involved appear much more reasonable. E_1 is a simpler formulation than d_r and will therefore be used. The accordance measures used are *RMSE*, d_1 , and E_1 , as shown in Table 4.1.

Figure 4.2: Graph of ground surface temperature on Shrubby Hummocks at point NGO-DD-2035_ST03 in winter 2015 demonstrating a temperature pattern that results in large negative NSE values. The winter season is highlighted in green. In this case, the lowest NSE value is -5364.1 for the ERA5/2 model, and the highest is -177.2 for the GT-JRA55 model.

4.5 Confidence intervals

Confidence intervals are used to assign a level of certainty to estimates of the population mean (also denoted by μ) of each accordance measure's assessment of a model. Through the use of bootstrap methods, a confidence interval describing how the sample mean $\bar{x}$ varies about the population mean μ can be obtained. Within a ranking framework, this allows for assessment of the certainty assigned to each rank, and in the case of overlapping confidence intervals, identifies multiple rank positions each model could occupy. The bootstrap t intervals method used in this work is specifically recommended by Hesterberg (2015) as more accurate for small samples than the common combination of nonparametric bootstrapping and bootstrap percentile confidence intervals.

The bootstrap t process for a single model m using a single accordance measure f begins

by determining the mean accordance $\bar{x}$ of m 's predictions $p_{Y,t,m}$ over all observation $o_{Y,t,m}$ for time intervals t . The resulting accordance values, expressed as $\tau_{p_{Y,t,m}}^f$, are then randomly re-sampled with replacement to form 10,000 samples containing the same number of elements as $\tau_{p_{Y,t,m}}^f$. The sample means of these 10,000 samples is represented by $\bar{x}^*$. The bootstrap principle means that variations of $\bar{x}^*$ about $\bar{x}$ approximate variations of $\bar{x}$ about μ (Hesterberg, 2015). These variations are simply calculated as

$$\delta^* = \bar{x}^* - \bar{x} \quad (4.1)$$

for each $\bar{x}^*$. The bootstrap t method extends this procedure by calculating the t statistic for each bootstrap resample, denoted as t^* . The distribution of t^* is approximately the same as the distribution of t , but accounts for the peculiarities of the distribution present in the given dataset (Hesterberg, 2015). t^* is calculated for each sample as

$$t^* = \frac{\delta^*}{SE^*} \quad (4.2)$$

for each δ^* , where SE^* is the standard error of the mean of the bootstrap resample. Standard error may be calculated formulaically in this case, but can also be obtained through an iterated bootstrap approach in situations when formula standard errors are not available (Efron and Tibshirani, 1993). The general formula for mean standard error is

$$SE = \frac{\sigma}{\sqrt{n}} \quad (4.3)$$

where σ is the standard deviation of the estimator about the parameter of interest, and n is the size of the sample being evaluated. The 10,000 values of t^* are then sorted in ascending order. For a 95% confidence interval, the values at the 2.5th percentile ($t_{0.025}^*$, 250th position)

and the 97.5th percentile ($t_{0.975}^*$, 9,750th position) represent values which are used to form the upper and lower boundaries of the confidence interval respectively. This reversal of endpoints is intentional, though easily overlooked with standard confidence intervals due to symmetry (Hesterberg, 2015).

$$[\bar{x} - t_{0.975}^* \cdot SE, \bar{x} - t_{0.025}^* \cdot SE] \quad (4.4)$$

Chapter 5

Experimentation

5.1 Outline

A diagram outlining the three major experiments conducted is presented in Figure 5.2 in a simple data-flow diagram notation shown as Figure 5.1. A more detailed rationale and expected outcomes for each experiment are outlined in the remainder of this chapter. Direct linkages between these outcomes and their contribution towards fulfilling the objectives of this thesis are drawn.

Figure 5.1: Data flow diagram symbology used in diagrams of experimentation procedures

Figure 5.2: Diagram of experiments conducted. The experiments progressively subdivide the observation data and model results into increasingly temporally and spatially homogeneous groups to systematically test the consistency of ranks across various accordance measures with theoretically decreasing levels of model error

5.2 Rankings by annual accordance

In this first experiment, the ability of the models (M) to simulate observations was tested across all points for which there is at least a full year of data by comparing against observations using the three accordance measures selected in Section 4.4 (F). The goal of this experiment is to determine whether the choice of accordance measure affects the resulting choice of GST model when evaluating model performance on annual or greater temporal resolution over a dataset containing observations collected on varied ground types.

The first step was to subdivide the observations timeseries taken at 159 points into one-

Figure 5.3: Diagram representing subdivision of complete point timeseries into annual time-series. o_y is the observation at point y , and $o_{y,A}$ is the set of all single-year observations within o_y . Individual years within $o_{y,A}$ can be referenced as $o_{y,Ax}$, where x is the year number within the set.

year intervals to ensure evaluations of model accordance are measured across a complete and common set of atmospheric and seasonal conditions. Incomplete-year data were discarded. An annual start date of the first of September was chosen, as this date captures the start date of the most observations without passing the collection end-date, and therefore maximizes the number of complete years that can be included in the analysis. The resulting dataset has a GST time-series for each point-year, with points that have more than one complete year of observations being represented by multiple point-years, yielding a total of 276 observation point-years. Each observation time-series in this step is represented by:

$$o_{y,Ax}$$

where y is the point the observation represents, and Ax represents the x th year of annual data (A) within that point. A diagram representing this process is presented as Figure 5.3.

The second step was to calculate accordancess across all point-years and models:

$$\tau = F(o_{Y,A}, p_{Y,A,M}) \quad (5.1)$$

A diagram illustrating the application of an accordance measure to a single observation time-series and a corresponding prediction is included as Figure 5.4.

Figure 5.4: Diagram representing comparison of observation timeseries $o_{y,t}$ against prediction timeseries $p_{y,t,m}$ using accordance measure f , where y is a single point, t is the time interval being compared and m is the name or number of the model whose prediction is being compared to observations. $\tau_{p_{y,t,m}}^f$ is the resulting accordance value.

In this experiment, this process would be repeated over the six different model predictions for each time period t over all points y . When applied to all point-years within each point and model combination, a dataset of 276 τ values is generated for each of the six models within each accordance function f . A mean of these 276 τ values can be calculated for each model m , yielding a set of six mean-accordance values per accordance measure f :

$$\overline{\tau_{Y,A,m}^f} = \frac{\sum_{i=1}^n \tau_{p_{i,m}}^f}{n} \quad (5.2)$$

where n is a count of all valid point-years in $Y \cdot A$. A general version of this process is illustrated in Figure 5.5.

The resulting mean accordances for each model m can then be ranked from smallest to

Figure 5.5: Diagram representing the process of calculating mean accordance of all provided τ values by model using one accordance measure f . The provided τ values are separated by model and a mean is calculated over all points by time interval t (year, season) and model m . Annual evaluation produces one set of model mean accordance values, while seasonal evaluation produces one set per season for a total of four.

largest difference from the perfect model value as defined by accordance measure f :

$$\zeta_{Y,A}^f = Z_f \left(\overline{\tau_{Y,A,1}^f}, \overline{\tau_{Y,A,2}^f}, \dots, \overline{\tau_{Y,A,6}^f} \right) \quad (5.3)$$

This process is illustrated for the mean accordance of six models in Figure 5.6.

Figure 5.6: Diagram representing ranking of mean accords of the predictions made by six models against the entire observational dataset. The ranking process (Z_f) receives the accordance values for the six models and ranks them in order from smallest to largest difference from the perfect model value as defined by accordance measure f .

Individual rankings are inspected and compared for each accordance function f , and differences in rankings are noted and analyzed.

Table 5.1: Sample model rankings by accordance function.

f	Model		
	1	2	...
$RMSE$	$\zeta_{Y,A,1}^{RMSE}$	$\zeta_{Y,A,2}^{RMSE}$	
d_1	$\zeta_{Y,A,1}^{d_1}$	$\zeta_{Y,A,2}^{d_1}$	
E_1	$\zeta_{Y,A,1}^{E_1}$	$\zeta_{Y,A,2}^{E_1}$	

The order of operations for this procedure ensures that the final output is an integer rank for each of the six models within each accordance measure. This output can be described as the ranking of the model accordance means, answering the question “which is the best performing model” based on different measures of accordance with clear, easily interpretable values that encapsulate the entire dataset.

5.3 Rankings by seasonal accordance

A second experiment is conducted to evaluate the effect of temporal granularity on perceived model competence expressed by measures of accordance. Willmott et al. (2012) and Legates and McCabe (2013) assert that consideration of seasonality in the use of accordance measures provides a more accurate baseline for model comparison, and therefore better assessments of model skill than the use of those same measures over entire observational datasets. This experiment aims to examine the extent to which seasonality changes the rankings of each model over different accordance measures. Practically, this entails evaluating accordance at the seasonal level by further subdividing observations data on pre-defined dates that act as seasonal boundaries. The dates chosen are listed in Table 3.1, and were selected to line up with the annual analysis and thus maximize the number of complete seasons captured.

These seasonal time-periods are expressed as:

$$O_{y,Su}, O_{y,Fa}, O_{y,Wi}, O_{y,Sp}$$

where summer, fall, winter and spring are denoted by Su , Fa , Wi , and Sp respectively, and where each season denotes the set of all instances of that season for point y . The process of seasonal timeseries splitting is depicted in Figure 5.7.

Figure 5.7: Diagram representing subdivision of complete point timeseries into seasonal timeseries. o_y is the observation at point y , and $o_{y,Su}$, $o_{y,Fa}$, $o_{y,Wi}$, and $o_{y,Sp}$ are the sets of all seasonal summer, fall, winter and spring observations respectively within o_y .

The next step was the calculation of model accordances for all point-seasons. Similar to the process for the ranking of annual model accordance means, a mean accordance value for each model is calculated, however as the data are subdivided by season, a set of six model mean values are generated for each season, for a total of twenty-four values for each accordance function f using equation:

$$\overline{\tau_{Y,t,m}^f} = \frac{\sum_{i=1}^n \tau_{i,m}^f}{n} \quad (5.4)$$

where n is a count of all valid point-seasons in $Y \cdot t$. This means the process outlined by Figure 5.5 will be run four times with the accordance values used as input being a sea-

sonal subset each time. Resulting mean accordances for each model-season combination are ranked with the same process as the annual values, and differences in rankings of models between accordance measures within each season are noted and analyzed for each accordance function f .

Table 5.2: Sample model/season rankings by accordance function.

f	Model							
	1				2			
$RMSE$	$\tau_{Y,Su,1}^{RMSE}$	$\tau_{Y,Fa,1}^{RMSE}$	$\tau_{Y,Wi,1}^{RMSE}$	$\tau_{Y,Sp,1}^{RMSE}$	$\tau_{Y,Su,2}^{RMSE}$	$\tau_{Y,Fa,2}^{RMSE}$	$\tau_{Y,Wi,2}^{RMSE}$	$\tau_{Y,Sp,2}^{RMSE}$
d_1	$\tau_{Y,Su,1}^{d_1}$	$\tau_{Y,Fa,1}^{d_1}$	$\tau_{Y,Wi,1}^{d_1}$	$\tau_{Y,Sp,1}^{d_1}$	$\tau_{Y,Su,2}^{d_1}$	$\tau_{Y,Fa,2}^{d_1}$	$\tau_{Y,Wi,2}^{d_1}$	$\tau_{Y,Sp,2}^{d_1}$
E_1	$\tau_{Y,Su,1}^{E_1}$	$\tau_{Y,Fa,1}^{E_1}$	$\tau_{Y,Wi,1}^{E_1}$	$\tau_{Y,Sp,1}^{E_1}$	$\tau_{Y,Su,2}^{E_1}$	$\tau_{Y,Fa,2}^{E_1}$	$\tau_{Y,Wi,2}^{E_1}$	$\tau_{Y,Sp,2}^{E_1}$

5.4 Rankings by seasonal and ground type accordance

Subdivision of data by ground type should provide more accurate and precise seasonal accordance values for each ground type and help to more clearly identify ground types that are most problematic for the tested models. This decrease in overall error per season/ground type group also places additional emphasis on how each accordance measure evaluates error. The aim of the third experiment is to investigate whether this additional emphasis on the evaluation of error results in differences in model rankings per season/ground type group.

This experiment builds on the temporal subdivision of the previous experiment by further grouping observation points by ground type. Overall mean accordance values and resulting rankings are then calculated on the combination of model, season and ground composition. The first step of this process is subdivision of timeseries by season as detailed in Section 5.3. Each observation point's timeseries are then grouped along the ground types specified in Table 3.2. This process results in 10 groups within each season. The remainder of this third experiment is essentially a repetition of the second on these 10 seasonally subdivided

Figure 5.8: Diagram representing subdivision of observation points into ground type (v) groupings. Seasonal observation timeseries are grouped based on the ground type of the point o_y at which the observations were collected. Past this point, each group v is treated as an independent set of seasonally separated observation timeseries.

observation datasets, and yields a ranking of the six models for each season-ground type combination, for a total of 40 sets of model rankings per accordance measure.

Chapter 6

Results

6.1 Model evaluation based on annual accordance values

This section presents the results of evaluating models using annual datasets over three accordance measures, as described in Chapter 4. The mean accordance over all points is calculated for each model, and the resulting rankings are compared across accordance measures, while the individual point-year accordance values are used to determine measures of confidence in each rank assigned.

The annual mean accordance values and ranks for all models over all measures are listed in Table 6.1. There are variations in the rankings within the three accordance measures indicating a difference in perceived model performance. While ranks remain stable for the four models at the bottom four ranks (GT-JRA55, GT-BSnow-ERA5, JRA55/2, ERA5/2, ranked 3 to 6 respectively), the top two models display a variation. GT-ERA5 and GT-BSnow-JRA55 are ranked 1, 2 by $RMSE$ and d_1 , and 2, 1 by E_1 .

95% confidence intervals were calculated for the means of each model/accordance measure using the means for individual point/years as the sample from which bootstrap resamples are drawn, and are listed in Tables C.1 to C.3. The confidence intervals, graphed as

Model	$RMSE$	d_1	E_1
GT-ERA5	1 / 2.8081	1 / 0.8173	2 / 0.5927
GT-JRA55	3 / 2.9088	3 / 0.8039	3 / 0.5919
GT-BSnow-ERA5	4 / 4.0699	4 / 0.7683	4 / 0.3909
GT-BSnow-JRA55	2 / 2.8237	2 / 0.8138	1 / 0.6013
ERA5/2	6 / 5.4148	6 / 0.6718	6 / 0.1695
JRA55/2	5 / 5.2461	5 / 0.6723	5 / 0.1931

Table 6.1: Model annual accordance ranks and means as calculated by different accordance measures. The first number in each column represents the model rank within that accordance measure while the second value is its mean accordance over all points.

Figure 6.1: 95% confidence intervals of all models on annual data within each accordance measure. Circles represent the mean accordance per model, and the horizontal lines represent the extent of the confidence interval associated with that point. Colours are used to differentiate the models within the diagram. Due to the different range of each accordance measure, the x-axis scale magnitudes vary by diagram.

Figure 6.2: Plausible ranks for each model within the 95% confidence intervals by accordance measure as calculated using annual data. The circles indicates the rank assigned to each model and the error bars represents the range of plausible ranks for that model based on its overlap with the 95% confidence intervals of adjacent models. The different colours are each associated with a model to make the identification of model rank fluctuations more apparent.

Figure 6.1 show how small the differences in performance are between models, and how the resulting overlap within the top three best-performing models introduces uncertainty into a decision of what the best model is for this dataset. Figure 6.2 better illustrates this uncertainty by presenting a plot of the models by rank, as well as error bars that represent the other plausible ranks for each model based on overlap between confidence intervals. This uncertainty is most evident on the top three and bottom two models.

6.2 Model evaluation based on seasonal accordance values

This section presents the results of evaluating models after subdividing the annual datasets into four seasonal datasets over the three accordance measures, as described in Chapter 4.

The seasonal means for each model within each accordance measure were calculated. These data are summarized in Table 6.2 by the mean of accordsances over the six models for each season. Annual accordance means from the previous section and a mean of the seasonal values are also included for comparison. This chart shows that in general, models performed best in the fall season, and the worst in the winter, with spring and summer ranked 3, 4 by d_1 and E_1 , and 4,3 by $RMSE$. $RMSE$ also differs from the other two measures by having its mean accordance of all models over four seasons indicate a better model than its annual accordance value, and by having the smallest proportional difference between the annual and seasonal mean values.

Table 6.2: Mean of accordance means of all models over annual dataset compared with mean of accordance means of all models over seasonal datasets.

Accordance Measure	Annual	Seasonal Mean	Summer	Fall	Winter	Spring
$RMSE$	3.8786	3.5351	3.2425	2.4255	4.9270	3.5454
d_1	0.7579	0.55855	0.5317	0.6505	0.4069	0.6451
E_1	0.4232	-1.231325	-0.3467	-0.0925	-4.1710	-0.3151

There is notable variation in the rankings for each model amongst the seasons. An illustration of this variation is presented as Table 6.3, where the mean rank of each model/accordance measure combination and the stability of its ranks across seasons is represented. This table indicates that the majority of models do not perform well enough across all seasons to consistently remain top ranked. Indeed, the three top-rated models under each accordance measure changed rank at least three times each.

Table 6.4 summarizes Tables 4.3 to 4.6, found in the Appendix, for $RMSE$, d_1 and E_1 by

Table 6.3: Rank stability table for models over all seasons based on complete rankings from Tables 4.3 to 4.6. The numerical value is the mean of all ranks assigned to the model/accordance measure combination over all seasons. The colour indicates rank stability, with green, yellow, orange, and red indicating one to four different ranks held over four seasons.

Model	$RMSE$	d_1	E_1
GT-ERA5	2.5	2.25	2.75
GT-JRA55	3	2.75	2.75
GT-BSnow-ERA5	3.75	3	3.75
GT-BSnow-JRA55	2.75	2	2.75
ERA5/2	4.75	5.5	4.75
JRA55/2	4.25	5.5	4.25

showing ranked model performance by season derived using the mean of all ranks assigned to that model for that season by the three selected accordance measures, represented as:

$$\frac{\sum_{j=1}^n \zeta_{Y,t,m}^j}{n} \quad (6.1)$$

where n is a list of the accordance measures $[RMSE, d_1, E_1]$ and t and m represent each season and model respectively. Annual rankings using the same methodology are included for comparison against individual seasonal ranks. This table shows the range of model rankings by season, and by its comparison to the annual rank, illustrates the difference the consideration of seasonality can make in model selection.

As was done for the annual data in the first experiment, 95% confidence intervals were calculated for the mean accordance of each model/season combination within each accordance measure. These confidence intervals were converted to plots of the models arranged by rank with error bars indicating overlap in bootstrap confidence intervals between models. These plots are presented as Figures 6.3 to 6.6. A key point visualized by these figures is that there is at least one case per season where the choice of accordance measure results in

Table 6.4: Overall model/season rank and mean of ranks assigned per season by all accordance measures used. The table indicates model performance per season as an overall rank between all models as evaluated by the accordance measures. The first number indicates overall rank as determined by the second number, which for the annual and seasonal columns indicates the mean of the ranks assigned to that model within the indicated season by the three accordance measures used. (T) denotes ties between models.

Model	Annual	Seasonal Mean	Summer	Fall	Winter	Spring
GT-ERA5	1 (1.33)	1(T) (2.5)	4(T) (4)	1 (1)	3 (3)	2 (2)
GT-JRA55	3 (3)	3 (2.83)	6 (5)	2 (2)	1 (1.33)	3 (3)
GT-BSnow-ERA5	4 (4)	4 (3.5)	1(T) (2.33)	4 (4)	4 (4)	4 (3.67)
GT-BSnow-JRA55	2 (1.67)	1(T) (2.5)	4(T) (4)	3 (3)	2 (1.67)	1 (1.33)
ERA5/2	6 (6)	6 (5)	3 (3.33)	5 (5.33)	6 (6)	5 (5.33)
JRA55/2	5 (5)	5 (4.67)	1(T) (2.33)	6 (5.67)	5 (5)	6 (5.67)

variations of a model's ranking within that season.

Figure 6.3: Plausible model ranks for each model/accordance measure within the 95% confidence intervals for the summer season. The coloured circles indicates the rank assigned to each model, while their coloured error bars represents the range of plausible ranks for that model based on its overlap with the 95% confidence intervals of adjacent models.

Figure 6.4: Plausible model ranks for each model/accordance measure within the 95% confidence intervals for the fall season. The coloured circles indicates the rank assigned to each model, while their coloured error bars represents the range of plausible ranks for that model based on its overlap with the 95% confidence intervals of adjacent models.

Figure 6.5: Plausible model ranks for each model/accordance measure within the 95% confidence intervals for the winter season. The coloured circles indicates the rank assigned to each model, while their coloured error bars represents the range of plausible ranks for that model based on its overlap with the 95% confidence intervals of adjacent models.

Figure 6.6: Plausible model ranks for each model/accordance measure within the 95% confidence intervals for the spring season. The coloured circles indicates the rank assigned to each model, while their coloured error bars represents the range of plausible ranks for that model based on its overlap with the 95% confidence intervals of adjacent models.

6.3 Model evaluation based on seasonal performance over different ground types

Further to Section 6.2, this section presents the results of evaluating datasets subdivided by season and by ground type over the three accordance measures, as described in Chapter 4.

The model mean accordsances of each combination of season and ground type were calculated for the three accordance measures. The Figures from 6.7 to 6.9 present the resulting season/ground type model accordance means, along with the seasonal model accordance means calculated in Section 6.2. As expected, model accordance means differ by the ground type being simulated, and these ground type model accordance means scatter about each model's seasonal accordance mean.

Figure 6.7: Graph of the mean accordsances of each ground type output by RMSE for each model over the four seasons.

The rankings of the mean model accordsances for each ground type and accordance measure are represented in Tables 6.5 to 6.8 by season. While each season contains variation

Figure 6.8: Graph of the mean accordances of each ground type output by d_1 for each model over the four seasons.

Figure 6.9: Graph of the mean accordances of each ground type output by E_1 for each model over the four seasons.

between $RMSE$, d_1 , and E_1 values within single terrain/model combinations, the most variation in terms of the size of rank discrepancy between $RMSE$, d_1 and E_1 values happens in the summer, where IWP, SG, SV, SH, and TV have at least two models with a rank difference of at least 4 rank positions, with IWP and SV each having a 5 position discrepancy on the ERA5/2 model. Spring and fall each have only one 4 position discrepancy, though spring has three 3 position discrepancies against fall's one such discrepancy. No terrain/model has a discrepancy greater than 2 positions over the winter.

95% confidence intervals were used to calculate the plausible ranks for each model at a season/terrain combination, graphs for which can be found in Appendix E. Summer values align with the previous findings of rank uncertainty among the models in Table 6.5, with three of the ten ground type graphs showing complete rank uncertainty, and only six of the thirty ground type/accordance measure combinations showing a narrow rank range for JRA55/2 and ERA5/2. The results for the fall and winter seasons show consistently narrower intervals than summer, however, neither shows a marked improvement over the plausible rankings calculated for the seasons over all ground types, particularly at the top three ranked models. The plausible ranks for spring show more variation among the ground types than the previous seasons, but all ground types still fail to consistently differentiate the top three models.

Validation objectives may differ on what an acceptable level of confidence for a model rank is, so confidence levels of 70% and 80% were also tested for some ground types and are graphed as Figure 6.10. In the tested cases, the 80% confidence interval improves the certainty of rankings for the bottom three-ranked models and in some cases, the certainty among the top three models compared to the 95% confidence interval. The 70% confidence interval displays some improvement in certainty over the 80% but not as consistently across ground types and accordance measures as the 80% displayed over the 95%.

Table 6.5: Ranks of models by ground type in summer as generated by $RMSE / d_1 / E_1$. Ground types are E: Eskers, IWP: Ice wedge polygons, IWT: Ice wedge troughs, SF: Sedge fen, SG: Sedges and grass, SV: Short vegetation, SH: Shrubby hummocks, SD: Snowdrifts, TS: Tall shrubs, TV: Till veneer.

Model	E	IWP	IWT	SF	SG	SV	SH	SD	TS	TV
GT-ERA5	4/4/4	4/4/4	1/1/1	1/2/1	4/4/4	4/5/5	4/4/4	1/3/1	1/1/1	4/2/4
GT-JRA55	3/2/3	6/3/6	3/3/3	2/1/2	6/3/6	6/3/6	6/3/6	4/2/4	3/3/3	6/4/6
GT-BSnow-ERA5	2/3/1	3/1/3	2/2/2	3/4/4	3/1/3	3/2/3	3/1/3	1/3/1	2/2/2	3/1/3
GT-BSnow-JRA55	1/1/2	5/2/5	4/4/4	4/3/3	5/2/5	5/1/4	5/2/5	3/1/3	4/4/4	5/3/5
ERA5/2	6/6/6	1/6/1	6/6/6	6/6/6	1/5/1	1/6/1	2/6/2	6/6/6	6/6/6	2/6/2
JRA55/2	5/5/5	2/5/2	5/5/5	5/5/5	2/6/2	2/4/2	1/5/1	5/5/5	5/5/5	1/5/1

Table 6.6: Ranks of models by ground type in fall as generated by $RMSE / d_1 / E_1$. Ground types are E: Eskers, IWP: Ice wedge polygons, IWT: Ice wedge troughs, SF: Sedge fen, SG: Sedges and grass, SV: Short vegetation, SH: Shrubby hummocks, SD: Snowdrifts, TS: Tall shrubs, TV: Till veneer.

Model	E	IWP	IWT	SF	SG	SV	SH	SD	TS	TV
GT-ERA5	1/1/2	1/1/1	1/1/1	1/1/1	1/1/1	3/3/3	1/1/1	1/1/1	1/2/1	1/1/1
GT-JRA55	6/4/5	3/3/2	2/3/2	3/3/2	2/2/2	1/1/1	2/2/2	4/4/3	4/4/3	2/2/2
GT-BSnow-ERA5	2/2/4	4/4/4	4/2/4	2/2/4	4/4/4	4/4/4	4/4/4	1/1/1	2/1/4	4/4/4
GT-BSnow-JRA55	5/3/6	2/2/3	3/4/3	4/4/3	3/3/3	2/2/2	3/3/3	3/3/4	3/3/2	3/3/3
ERA5/2	3/5/1	6/5/6	6/5/5	6/5/6	6/5/6	6/5/6	6/5/5	6/5/6	6/5/5	6/5/5
JRA55/2	4/6/3	5/6/5	5/6/6	5/6/5	5/6/5	5/6/5	5/6/6	5/6/5	5/6/6	5/6/6

Figure 6.10: Plausible model ranks for Sedges and Grass in spring at 95%, 80% and 70% confidence intervals.

Table 6.7: Ranks of models by ground type in winter as generated by RMSE / d_1 / E_1 . Ground types are E: Eskers, IWP: Ice wedge polygons, IWT: Ice wedge troughs, SF: Sedge fen, SG: Sedges and grass, SV: Short vegetation, SH: Shrubby hummocks, SD: Snowdrifts, TS: Tall shrubs, TV: Till veneer.

Model	E	IWP	IWT	SF	SG	SV	SH	SD	TS	TV
GT-ERA5	1/1/1	2/1/2	3/3/3	3/2/3	3/3/3	3/3/2	3/3/3	3/3/3	3/1/3	3/3/3
GT-JRA55	2/2/2	3/3/3	2/2/1	2/3/2	1/1/1	2/2/3	1/1/1	1/1/1	1/2/1	1/2/1
GT-BSnow-ERA5	4/4/4	4/4/4	4/4/4	4/4/4	5/4/6	6/6/6	4/4/4	3/3/3	4/4/4	6/4/6
GT-BSnow-JRA55	3/3/3	1/2/1	1/1/2	1/1/1	2/2/2	1/1/1	2/2/2	2/2/2	2/3/2	2/1/2
ERA5/2	5/5/5	6/6/6	6/6/6	6/6/6	6/6/5	5/5/5	6/6/6	6/6/6	6/6/6	5/6/5
JRA55/2	6/6/6	5/5/5	5/5/5	5/5/5	4/5/4	4/4/4	5/5/5	5/5/5	5/5/5	4/5/4

Table 6.8: Ranks of models by ground type in spring as generated by $RMSE / d_1 / E_1$. Ground types are *E*: Eskers, *IWP*: Ice wedge polygons, *IWT*: Ice wedge troughs, *SF*: Sedge fen, *SG*: Sedges and grass, *SV*: Short vegetation, *SH*: Shrubby hummocks, *SD*: Snowdrifts, *TS*: Tall shrubs, *TV*: Till veneer.

Model	E	IWP	IWT	SF	SG	SV	SH	SD	TS	TV
GT-ERA5	4/4/4	1/2/1	1/2/1	3/2/3	1/1/1	2/2/2	1/1/3	3/3/3	3/1/3	2/2/3
GT-JRA55	3/3/2	4/5/3	4/6/3	2/3/2	3/4/3	3/4/3	3/3/2	1/1/1	2/4/1	3/4/2
GT-BSnow-ERA5	2/2/3	3/1/4	3/1/4	6/6/6	4/3/4	4/3/4	4/4/4	3/3/3	6/6/6	4/3/6
GT-BSnow-JRA55	1/1/1	2/3/2	2/3/2	1/1/1	2/2/2	1/1/1	2/2/1	2/2/2	1/5/2	1/1/1
ERA5/2	5/5/5	5/4/5	5/4/5	5/5/5	5/5/5	5/5/5	6/5/6	6/6/6	5/3/5	5/5/4
JRA55/2	6/6/6	6/6/6	6/5/6	4/4/4	6/6/6	6/6/6	5/6/5	5/5/5	4/2/4	6/6/5

Chapter 7

Interpretation

In this chapter, the results of the experiments in Chapter 6 are interpreted in the context of the three main research objectives.

7.1 Objective 1: The effect of using different measures of accordance in evaluating model competence in GST pre- diction using annual data

The results of this experiment are not far from the expectation of small variations in ranks between accordance measures, with the only variation being at the top two ranks where the GT-ERA5 and GT-BSnow-JRA55 models alternated as the top two models, usually by the second smallest margin between models within each accordance measure.

A notable result of experimentation on annual data is the performance ranking of the GT-BSnow-ERA5 model. The GT-BSnow-ERA5 and GT-BSnow-JRA55 models integrate the additional complexity of GEOTop’s “blowing snow” module, with the only difference between these models being the reanalysis data driving them. The rankings at Table 6.1 show that the

GT-BSnow-JRA55 model displays the second-best overall performance. GT-BSnow-ERA5 is ranked fourth among all measures by a far wider margin than the differences between the top three models. On the surface, this discrepancy would appear to indicate a difference in quality between ERA5 and JRA-55 data, however, the rankings of GT-ERA5 and GT-JRA55 as first and third respectively, a reversal of the order of GT-BSnow-JRA55 and GT-BSnow-ERA5, indicates that it is not a difference in data quality that causes this ranking discrepancy, but rather it is more likely a result of the interplay between model and data. This reinforces the notion that apart from the choice of accordance measure, the interactions between data patterns in the driving data and the model processes also affect perceived model competence. Any attempt at determinations of model competence should therefore consider this complete stack of inputs and processes used as part of the evaluation framework.

Overall, the amount of rank variation between different accordance measures in this experiment does not by itself make a strong case that the choice of accordance measure has a substantial effect on model selection when annual data are used to assess model competence. This is likely due to the relatively large timescale over which the effects of prediction errors are distributed, resulting in more favourable and stable evaluations of model performance. The 95% confidence intervals calculated for the means of each model, when translated to reflect the rank and plausible ranks of each model in Figure 6.2 also displays enough uncertainty to make the top three models essentially interchangeable.

7.2 Objective 2: The effect of temporal subdivision of datasets into seasons on evaluations of model competence in GST prediction

The expectation for this experiment was that differences in how a single model is ranked across different seasons by the same accordance measure and in the order models are ranked by different accordance measures within the same season would be found. Findings of the first difference are best summarized in Table 6.3, which illustrates the overall skill and the number of changes in rank of each model over all seasons. In brief, this table shows that each model does not achieve a uniform rank across all seasons, and in some cases, achieves a different rank for each season. The overall skill of all models within each season is illustrated in Table D.1, where the means of all model accordance means $\left(\overline{\tau_{PY,Su,M}^f}, \overline{\tau_{PY,Fa,M}^f}, \overline{\tau_{PY,Wi,M}^f}, \overline{\tau_{PY,Sp,M}^f} \right)$ are calculated for each season. An unexpected aspect of the rankings is that the underlying values generated by E_1 for each observation/model comparison, and the resulting seasonal model means (Table D.1) indicate that most model predictions in the dataset have less predictive reliability ($E_1 < 0$) than the observational mean for each seasonal time-series. This is also true for accordance measures related to E_1 ($NSE < 0$, $d_r < 0.5$). These seasonal values contrast strongly against the annual accordance values for these measures, also contained in Table D.1, which indicate that all models are better predictors than the observational mean. The annual values for $RMSE$ and d_1 show more agreement with their respective seasonal values. An explanation of this discrepancy is that the NSE and its derivatives measure how good a model is in comparison to a benchmark model, in this case the observational mean. The observational mean can be a poor predictor for strongly seasonal timeseries, and a good predictor for timeseries that are fluctuations around a relatively constant mean (Schaepli and Gupta, 2007). Some timeseries within the

dataset match the description of fluctuations around a relatively constant mean, but many are strongly seasonal, making the observational mean largely a poor predictor for annual timeseries, resulting in higher accordance values for NSE . When using seasonal timeseries, the observational mean is more localized and a better predictor, requiring model predictions to be more accurate than annual predictions to achieve a similar score. The smaller timescale of seasonal datasets also magnifies the influence of errors due their effect being spread out over only a season instead of an entire year. This limitation should borne in mind in validation methods that employs NSE , d_r or E_1 .

Table 6.2 shows that for all selected measures applied to seasonal data, winter predictions achieve the poorest mean accordance, and fall predictions achieve the best mean accordance. The difficulty in simulating winter is predictable due to the added complexity that snow cover and underground water/ice dynamics entail. Potential reasons for the models' success in simulating fall are that fall is the most linear season, with a downward trend in temperature, often followed by the start of the zero curtain, both of which are reasonably well simulated. This contrasts with summer, whose temperature sees a gradual rise, a peak, and the start of a gradual fall. The effects of snow cover also do not usually impact fall as much as winter and the start of spring, making simulation less complicated. For the dimensionless accordance measures, the annual timeseries achieved better mean scores than any individual season's mean. This was not the case for $RMSE$, with its fall mean values showing better performance than the annual timeseries' mean. A possible reason for these differences is the nature of what these measures quantify. $RMSE$ simply reflects the error between the observation and prediction while d_1 and E_1 reflect comparisons of error sums against benchmark models that involve the observational mean, a benchmark that fits each individual season poorly when evaluating annual datasets as noted earlier in this section. When this benchmark is localized to seasonal datasets, the increased ratio of prediction error to benchmark error results in lower model scores in d_1 and E_1 . This localization does not affect $RMSE$ as

it lacks this benchmark aspect.

In terms of the rankings displayed in Table 6.4, a notable anomaly is the high rankings assigned to the ERA5/2 and JRA55/2 models for the summer season, while these same models alternate the bottom two ranks for all other seasons and on annual data. This points to the additional complexity introduced into the model to better simulate seasonal effects related to cold weather dynamics and the energy transfer mechanisms between the atmosphere and the ground, and how in some simple cases, this complexity may be a liability. This anomaly also serves to demonstrate the second expectation for this experiment, in that ERA5/2 and JRA55/2's ranks change depending on the accordance measure for summer and fall. Model ranks for spring show the most fluctuation, with no model's rank being unanimously agreed upon by all accordance measures.

All seasons display some degree of uncertainty in their model rankings, though this uncertainty is most evident in the summer (Figure 6.3), where relatively close means for the top four ranked models result in uncertainty between the top three ranks at best (E_1) and between all six ranks at worst ($RMSE$), even with small confidence intervals for all accordance measures. This once again illustrates the importance of understanding the accordance measures being used, even within a simplifying process like ranking. Knowledge of a measure's range and scaling is necessary in order to determine whether model rank interchangeability could be indicative of large variations in accordances within a season or simply of models that achieve very similar seasonal mean accordances due to an accordance measure's narrow range.

7.3 Objective 3: The effect of the consideration of ground characteristics on evaluations of model competence for GST prediction using seasonal datasets

An expected outcome of calculating accordance values on seasonal datasets subdivided by ground type is that model skill in predicting different ground types is quantified, and subsequently, ground types can be categorized by the difficulty of their simulation by season. Another expected outcome is that variations in rankings between accordance measures can be localized to specific ground type/season combinations for further investigation. Such rank differences may reinforce or refute the findings on relationships between accordance measures of Section 7.2.

The first expected outcome is addressed by Figures 6.7 to 6.9, where the poor accordance of Short Vegetation in summer and fall exemplify a ground type/season combination that is poorly simulated. Among these poor performers, some are notable for the disproportionate effect their magnitudes have on the overall seasonal accordance means. Tall shrubs and Snowdrifts achieve large negative model accordance mean values for all models in E_1 , far below those of other ground types in winter and spring. The magnitude of the values is due to E_1 's unbounded negative range. While the Tall Shrubs and Snowdrift model accordance means are just as out of step in d_1 , the bounded nature of that measure limits the influence these two ground types exert on the seasonal mean. The numerical spread of ground type/season accordance means indicates that a model rating over one ground type is not transferable to another, or even to the same ground type in a different season. This variability in model skill based on ground type and season also challenges the practice of inferring model validity over an area based on point evaluations. The potential for some season and ground type combinations to produce extreme accordance values further highlights the importance

of making conscious accordance measure selections so that the difference between reactions of accordance measures to specific data patterns and legitimately poor model performance can be recognized.

The second expected outcome can be observed in Tables 6.5 to 6.8. As previously noted, substantial rank difference is found in summer over the Ice Wedge Polygon, Sedges and Grass, Shrubby Hummocks, Short Vegetation and Till Veneer ground types. It is primarily the rank assigned by d_1 that creates this discrepancy, with $RMSE$ and E_1 showing complete rank agreement on all but the Sedge Fen and Short Vegetation ground types. A look at Figure E.1 shows that the plausible ranks for Ice Wedge Polygon, Sedges and Grass, Shrubby Hummocks, Short Vegetation and Till Veneer display complete rank uncertainty using d_1 . $RMSE$ and E_1 offer more certainty for some of these ground types by being mostly matched in model ranks and producing similar plausible rank patterns. d_1 's disagreement with $RMSE$ and E_1 appears exclusively on ground types for which $RMSE$ and E_1 rate the simple ERA5/2 and JRA55/2 models as the two best, while d_1 ranks them as the worst. This once again demonstrates the potential for very specific circumstances along with accordance measure selection to have a large impact on determinations of model competence. The difference in rank certainty in these ground types can largely be attributed to d_1 's small range (0 to 1), which would result in even a small confidence interval encompassing all model means. Rank differences over the remaining three seasons, though not as dramatic as those over the summer allow for the identification of Esker in the fall and Ice Wedge Polygon, Ice Wedge Trough, and Tall Shrubs in the spring as the ground type/season combinations that cause the most disagreement amongst accordance measures. Winter displays only minimal rank disagreement amongst those same accordance measures.

Chapter 8

Discussion

8.1 Rank variations and their causes

The aim of this work was to highlight the influence of accordance measure choice on determinations of model validity for a particular purpose. The results obtained in Chapter 6 demonstrate this influence and the resulting changes in rank by model. Predominantly, there is agreement on ranks among the accordance measures, however, further analysis in Chapter 7 indicates that rank differences are the result of a more complicated interplay between all inputs and processes. This is evident in Section 7.1 where the GT-ERA5 model ranked higher than GT-JRA55 model, while the roles were reversed for the GT-BSnow-ERA5 and GT-BSnow-JRA55 models, indicating that patterns within the nominally similar ERA5 and JRA-55 reanalysis data affect each model in different ways. The discrepancy between E_1 's seasonal and annual means first shown in Section 6.2 and further explored in Section 7.2 is another such example. While it is not inaccurate to say that accordance measures causes rank differences, a more accurate statement would be that rank differences result from interplays within the complete “stack” of inputs and processes being tested, in this case model driving data, model, observations and accordance measure. Each choice in this stack mat-

ters and should be carefully chosen to suit the objective of the modelling exercise. Testing should therefore be conducted and results reported for complete stacks, not just generalized assessments of the performance of their individual elements.

8.2 The effect of more specific data subsets on rank variations

As the data intervals used were progressively constrained from annual to seasonal cycles, and finally subdivided spatially, minor rank differences seen in annual data increased on seasonal data, and spatial subdivision revealed the relative influence of season/ground type combinations on the aggregated annual and seasonal evaluations. This progression and the differences noted from each stage to the next reveal a problem with assessments made based on data that are not constrained more granularly than annual intervals: relevant details of stack performance can be drowned out within a larger, more general dataset. The values of E_1 in Chapter 6 present a good example of the problem. The annual E_1 accordance value at Table 6.2 indicates that model performance is better than the observed mean as a predictor. The seasonal values in the same table indicate otherwise, with the value for winter indicating significantly worse performance. Inspection of Figure 6.9 reveals that the large seasonal negative values can largely be credited to the outsized influence of two or three ground types, especially in spring and winter. Subsection 2.3.6 shows that these large negative values can be attributed to the interaction of specific seasonal ground effects with the NSE accordance measure and its derivatives. Without assessing data at more granular levels than annual, these effects exert disproportionate influence while remaining hidden. It is therefore important to understand the potential for this type of error in each stack, to make informed decisions on what shortcomings of a stack are acceptable, and to test for those that are not

acceptable. Awareness of these shortcomings will affect how confidently model predictions at different locations can be interpreted.

8.3 Reproducibility

The argument for a "whole stack" approach to model testing highlights the perceived differences in model competence that can be attributed individually to model driving data, data granularity and accordance measures used, as well as those attributed to the response of each aspect to all others in the stack. An implication of this view is that the integrity of the stack must be maintained in subsequent model tests for the results to be comparable. Reproduction of this type of testing therefore requires that not just the process be made public, but the data used be provided for scrutiny and future tests. This is the only way to ensure that the specific interactions between data and process that shape the results are accurately reproduced.

8.4 The value of interpretability

Section 2.3.3 discussed the interpretability of some individual accordance measures. One of the oft-cited benefits of dimensioned measures of accordance is that they are easily interpreted due to their results being expressed in the same units as those of the values being compared. Willmott and Matsuura (2005) noted that for the most commonly reported dimensioned measure of error, $RMSE$, this notion of interpretability is misleading as there is no consistent functional relationship between $RMSE$ and average error. Dimensionless measures of accordance on the other hand present a scale limited on at least the positive range, with important variations being on scaling through the range and on the point at which a model's performance represents an improvement on the benchmark it is being compared against, often the observational mean. There is practical value to both approaches in model

testing, in that either approach can be used to quantify both baseline model performance and improvements in subsequent model versions. This thesis was formulated on the idea that the true metric of interest when selecting a model is not individual model accordance values, but rather model performance relative to other models. This metric can easily be expressed as rankings, a universal, easily interpretable, unidirectional scale. In addition to the ranks, confidence intervals are also calculated on the accordance measures and presented as the potential ranks that each model could occupy within the group in which it is being compared. With the data used for this thesis, potential ranks often indicate that the top three models are interchangeable, and in some cases, all ranks are interchangeable. This pattern can be indicative of various issues including very similar performance from different models, large variation of accordance values for a model, or an overall lack of data to reduce the size of the confidence intervals. There is value in using the uncertainty and underlying limitations of the rankings presented to guide final model selection, or ideally, identifying the steps necessary to minimize model rank interchangeability prior to making a selection.

8.5 Implications of findings on permafrost research

This research has revealed some peculiarities of assessing model performance in a permafrost environment. These findings include the limitations of NSE and its derivatives in temperature regimes that hover about 0°C, and the large variation in the ranking of simple models in cold versus hot temperature regimes. Gubler et al. (2011) demonstrated that even over homogeneous terrain less than 14 m apart, notable variations of mean annual ground surface temperature can be found. Given this variation within the same ground type, and the variations that exist in permafrost types and locations, there are likely to be many more ground effects that when combined with elements of a model evaluation stack will yield unexpected results. This point speaks to the problem with current validation practices discussed in Sub-

section 2.2.2. Namely, generalizations of model performance over specific ground types should only be done very carefully over homogeneous ground data points, with a process in place for identifying and analysing anomalous results within the data. Generalizing model performance over non-homogeneous areas using points of varied ground types should be avoided, or only carried out for objectives that recognize the many potential types of error that can be introduced by each element in the stack. The points may be representative of the ground in aggregate, but part of the model's evaluation is based on how each individual point interacts with the stack.

8.6 Selecting an accordance measure for GST

This thesis has demonstrated that accordance measure selection affects how models for the simulation of GST are ranked. As objectives of modelling differ, there are no rules for the selection of a most suitable accordance measure. A selection process should take into account the characteristics and interactions particular to the evaluated dataset. In the case of the annual data used in this work, d_1 and $RMSE$ achieved identical ranks for all models, with E_1 differing only on the top-two-ranked models. The cyclical nature of annual data diminishes the effectiveness of a benchmark to measure model performance against, as used by d_1 and E_1 . Given the additional complication that this step introduces with no additional benefit, $RMSE$ is a more appealing option for this case. Seasonal evaluations of this particular data series show d_1 to be the most suitable choice. d_1 is not subject to the problems exhibited by NSE -based accordance measures at temperatures close to 0°C , as detailed in Section 2.3.6. Its constrained negative range also serves to limit the influence of extreme values within the dataset. d_1 's rankings of the simple models ERA/2 and JRA/2 as the two bottom models remains consistent through all seasons, and it also achieves relatively consistent plausible ranks through all seasons.

8.7 Assumptions and further uncertainties

This thesis highlighted some cases where the cause of different ranks being assigned to a model by different accordance measures was not fully understood and was attributed to the underlying error patterns between the observation and simulation data. As a result, the reaction of accordance measures to different data and error patterns generated by each season and ground type was a topic of speculation throughout this work. A better understanding of the strengths and limitations of each measure in evaluating model skill would be gained by further studying the cause of these reactions and the specific effects they have on the accordance assigned. Investigations in how these reactions and effects can be mitigated by accordance measures, such as assigning weights to values to balance the influence of different ground types within the dataset could be useful in furthering the applicability of these measures.

Aspects of this work dealing with ground classification were based on work done by Cao et al. (2019). It was assumed that for the purposes of this work, the ground classifications were sufficient to capture the full spectrum of ground types represented within the dataset, and that differences in observations within a group were based on the state of the ground at the time the observation was made, not on differences in composition due to misclassification or changes in ground cover over time. The further subdivision of the ground type groupings based on additional variables, such as slope angle and slope aspect, and whether this results in less variability and more confident determinations of model rank by group is a potential area of further research.

A decision was made that for the purpose of maximizing the amount of useable data within the dataset, observation years would be based on a start date of September 1st, and subsequently, seasons were aligned to each be three months long with fall beginning on its meteorological start date of September 1st instead of its astronomical date of September 21st. The effect of seasonality determined by data patterns or astronomical season start

dates, as well as more granular subdivision of observations to isolate certain data patterns may bolster a study of the response of accordance measures to different data patterns as suggested previously in this section.

As previously noted in this work, data availability is a constant limitation of any study involving observations in Arctic and sub-Arctic environments. As more data become available over time, further study should be conducted on the rate at which additional data narrow determinations of model ranks and potential ranks.

Chapter 9

Conclusion

The aim of this work was to investigate the difference that the choice of accordance measure makes on the inferred relative validity of a model for a particular purpose. This was done by investigating the differences in model rankings across various accordance measures, and highlighting features of the accordance measures and patterns over the dataset that may be the cause of these differences. The results have demonstrated that there are indeed differences in how each accordance measure evaluates and ultimately ranks models, and while the magnitude of these differences reflects the reaction of the accordance measure formulation to data patterns, choices made on subdivision and grouping of the data also affect the outcome.

9.1 The effect of different measures of accordance on determinations of model competence in the prediction of GST using annual data

The first objective of this thesis was to evaluate the effect of different measures of accordance on determinations of model competence in the prediction of GST. When three different ac-

cordance measures are applied to timeseries representing a full year, the rankings assigned to the six models show only minor variation in E_1 's ranking of the top two models. As mean accordances over the entire dataset are close for the top three models, confidence intervals converted to potential ranks show uncertainty in the ranking of these top three models indicating that any of them have the potential to be the model that best approximates observations under the measure used.

A broad conclusion of this objective pertains to the importance of evaluating model performance at a scale suitable to the heterogeneity present in the data. In strongly seasonal datasets, even large differences between model reactions to seasonal effects may be normalized by a large number of data points in which the models produce similar reactions. This effect is reflected in the largely identical rankings obtained with the use of annual data.

9.2 The effect of seasonality on rankings of model accordance across accordance measures

The second objective of this thesis was to evaluate the effect of seasonality in rankings of model accordance and how this varies by accordance measure. Seasonal differences in how accordance measures rank models were found, with no one model remaining top-ranked through all four seasons under any accordance measure. Rank differences could partly be attributed to d_1 and E_1 's evaluations of models against a benchmark that becomes more precise as the fluctuations and timescale in a dataset are reduced, a quality that $RMSE$ does not share. $RMSE$ values are therefore not greatly affected by this reduction in scale, while d_1 and E_1 values assign models lower values due to their comparison against a better benchmark. The top ranking models continued to yield accordance means close to each other over most seasons, which resulted in the confidence intervals encompassing the three

top models in all but one case, once again resulting in rank uncertainty among the top three models.

The variations in model ranks between accordance measures that result from using seasonal data highlight the importance of having a well-defined objective to inform data and accordance measure selection. An understanding of the objective would reveal which data subsets and model characteristics are most pertinent and the data and accordance measure used to evaluate models could then be selected accordingly. As demonstrated in Section 8.6, relating this understanding to an objective to be achieved can guide a reasoned model selection process.

9.3 The effects of spatial subdivision by grouping ground types on seasonal measures of accordance

The third objective evaluates the effect that grouping data points by ground type within the seasonal datasets of the second objective has on model ranks. Study of the mean accords of each ground type by season and accordance measure found specific combinations of ground type and season that were poorly simulated by the models, in some cases so poorly as to produce anomalous values within the dataset that would have strongly affected the seasonal mean of all ground types. This finding demonstrates the importance of the additional context for model performance that consideration of the granularity of data provides. It also highlights a problem inherent in generalizations of model skill from one location to another, especially in cases where the point composition is not considered, or an inference is being made from a point to an area. The problem of different compositions and relative influence of each ground type can only be addressed by precise knowledge of the composition of the destination for which extrapolation is desired. For this dataset, the uncertainty among the

top three models when using 95% confidence intervals remained with few exceptions, once again demonstrating the uncertainty in selecting a best model with limited data, even when efforts are made to reduce variation within each group of accordance values.

The specific data/model interactions observed in the previous two objectives and their effect on model ranks illustrates another broad conclusion of this work. In ranking a model, the reanalysis model driving data, observation data, and accordance measure used each contribute to the accordance value a model achieves, and therefore its ranking. Consideration of each combination of these elements as a “stack” matched to an objective with which models can be ranked would enable more consistent comparisons of model ability, as alluded to in Section 7.1.

9.4 Final thoughts

The unique effects that interactions between model driving data, data subdivision criteria, and accordance measure used have on model rankings were key findings of this work. If the model driving data, observations data and evaluation processes are considered a “stack” that contains the ingredients for these interactions, and where the model predictions are the only interchangeable components, then comparable evaluations of the performance of various models can be made using the same stack.

In selecting the stack elements to be used to evaluate a model, consideration must be given to the objective of the model’s employment. Understanding a dataset’s spatial and temporal extents, the strengths and limitations of the accordance measures being used, and tailoring or selecting these elements based on the needs of a well-defined objective will enable the creation of a stack of components that identifies the model that best achieves that objective. This strong link between an objective and the stack used to determine the best model may also have the beneficial side-effect of discouraging generalizations of model

competence to spatial and temporal circumstances that differ greatly from those represented by the stack.

Ranks were used in this work to simplify the interpretation of accordance measure results by providing a clear representation of which models perform best. The addition of plausible ranks based on the 95% confidence intervals provide a simple representation of the uncertainty of the model selection process, and create a connection to the limitations of the chosen accordance measures and data. It is hoped that simple portrayals of this information enhance understanding of the challenges in evaluating and selecting models, and in light of shortcomings in rank certainty that must often be accepted, highlights the importance of specific objectives to validate the chosen model against to ensure that it meets a minimum standard.

References

- Anderson, E. A. (1976). A point energy and mass balance model of a snow cover. Technical report, Office of Hydrology, National Weather Service, Silver Spring, MA, USA.
- Badache, M., Eslami-Nejad, P., Ouzzane, M., Aidoun, Z., and Lamarche, L. (2016). A new modeling approach for improved ground temperature profile determination. *Renewable Energy*, 85:436–444, DOI: 10.1016/j.renene.2015.06.020.
- Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A. A., Fath, B. D., and Andreassian, V. (2013). Characterising performance of environmental models. *Environmental Modelling and Software*, 40:1–20, DOI: 10.1016/j.envsoft.2012.09.011.
- Bryant, B. P., Hamel, P., and Mandle, L. (2018). Look beyond peer-reviewed literature and traditional validation when assessing ecosystem services modeling efforts: A response to Ochoa and Urbina-Cardona’s review. *Ecosystem Services*, 30:1–2, DOI: 10.1016/j.ecoser.2018.01.004.
- Cao, B., Quan, X., Brown, N., Stewart-Jones, E., and Gruber, S. (2019). GlobSim (v1.0): Deriving meteorological time series for point locations from multiple global reanalyses. *Geoscientific Model Development*, 12(11):4661–4679, DOI: 10.5194/gmd-12-4661-2019.
- Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250, DOI: 10.5194/gmd-7-1247-2014.
- Chen, X., Yuan, H., and Xue, M. (2018). Spatial spread-skill relationship in terms of agreement scales for precipitation forecasts in a convection-allowing ensemble. *Quarterly Journal of the Royal Meteorological Society*, 144(710):85–98, DOI: 10.1002/qj.3186.
- Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T. (2018). The international land model benchmarking (IL-AMB) system: design, theory, and implementation. *Journal of Advances in Modeling Earth Systems*, 10(11):2731–2754, DOI: 10.1029/2018MS001354.

- Council for Regulatory Environmental Modeling (2009). Guidance on the development, evaluation, and application of environmental models. *USEPA Publication*, (March):90.
- de Grandpré, I., Fortier, D., and Stephani, E. (2012). Degradation of permafrost beneath a road embankment enhanced by heat advected in groundwater. *Canadian Journal of Earth Sciences*, 49(8):953–962, DOI: 10.1139/E2012-018.
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597, DOI: 10.1002/qj.828.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman Hall, New York, DOI: 10.1007/978-1-4899-4541-9.
- Eker, S., Rovenskaya, E., Obersteiner, M., and Langan, S. (2018). Practice and perspectives in the validation of resource management models. *Nature Communications*, 9(1):5359, DOI: 10.1038/s41467-018-07811-9.
- Endrizzi, S., Gruber, S., Dall’Amico, M., and Rigon, R. (2014). GEOTop 2.0: Simulating the combined energy and water balance at and below the land surface accounting for soil freezing, snow cover and terrain effects. *Geoscientific Model Development*, 7(6):2831–2857, DOI: 10.5194/gmd-7-2831-2014.
- Fox, R. A., Steel, R. G. D., and Torrie, J. H. (1961). Principles and procedures of statistics with special reference to the biological sciences. *The Incorporated Statistician*, 11(3):170, DOI: 10.2307/2987461.
- Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., da Silva, A. M., Gu, W., Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, DOI: 10.1175/JCLI-D-16-0758.1.
- Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z. L., and Zhang, M. (2011). The community climate system model version 4. *Journal of Climate*, 24(19):4973–4991, DOI: 10.1175/2011JCLI4083.1.

- Gisnas, K., Westermann, S., Vikhamar Schuler, T., Melvold, K., and Etzelmüller, B. (2016). Small-scale variation of snow in a regional permafrost model. *Cryosphere*, 10(3):1201–1215, DOI: 10.5194/tc-10-1201-2016.
- Gruber, S. (2012). Derivation and analysis of a high-resolution estimate of global permafrost zonation. *Cryosphere*, 6(1):221–233, DOI: 10.5194/tc-6-221-2012.
- Gruber, S., Brown, N., Stewart-Jones, E., Karunaratne, K., Riddick, J., Peart, C., Subedi, R., and Kokelj, S. (2018). Air and ground temperature, air humidity and site characterization data from the Canadian Shield tundra near Lac de Gras, Northwest Territories, Canada, v. 1.0 (2015-2017). DOI: 10.5885/45561XD-2C7AB3DCF3D24AD8.
- Gubler, S., Endrizzi, S., Gruber, S., and Purves, R. S. (2013). Sensitivities and uncertainties of modeled ground temperatures in mountain environments. *Geoscientific Model Development*, 6(4):1319–1336, DOI: 10.5194/gmd-6-1319-2013.
- Gubler, S., Fiddes, J., Keller, M., and Gruber, S. (2011). Scale-dependent measurement and analysis of ground surface temperature variability in alpine terrain. *Cryosphere*, 5(2):431–443, DOI: 10.5194/tc-5-431-2011.
- Guo, D., Wang, H., and Li, D. (2012). A projection of permafrost degradation on the Tibetan Plateau during the 21st century. *Journal of Geophysical Research Atmospheres*, 117(5):D05106, DOI: 10.1029/2011JD016545.
- Hersbach, H., Bell, B., Berrisford, P., Horányi, A., Sabater, J. M., Nicolas, J., Radu, R., Schepers, D., Simmons, A., Soci, C., and Dee, D. (2019). Global reanalysis: goodbye ERA-Interim, hello ERA5. *ECMWF Newsletter*, 159(Spring 2019):17–24, DOI: 10.21957/vf291hehd7.
- Hersbach, H. and Dick, L. (2016). ERA5 reanalysis is in production. *ECMWF Newsletter*, 147(147):7.
- Hesterberg, T. C. (2015). What teachers should know about the bootstrap: resampling in the undergraduate statistics curriculum. *American Statistician*, 69(4):371–386, DOI: 10.1080/00031305.2015.1089789.
- Hinzman, L. D., Goering, D. J., and Kane, D. L. (1998). A distributed thermal model for calculating soil temperature profiles and depth of thaw in permafrost regions. *Journal of Geophysical Research Atmospheres*, 103(D22):28975–28991, DOI: 10.1029/98JD01731.
- Hoffman, F. M., Koven, C. D., Keppel-Aleks, G., Lawrence, D. M., Riley, W. J., Randerson, J. T., Ahlström, A., Abramowitz, G., Baldocchi, D. D., Best, M., Bond-Lamberty, B., De Kauwe, M., Denning, A. S., Desai, A., Eyring, V., Fisher, R., Gleckler, P. J., Huang, M., Hugelius, G., Jain, A. K., Kiang, N. Y., Kim, H., Koster, R. D., Kumar, S. V., Li, H., Luo, Y., Mao, J., McDowell, N. G., Mishra, U., Moorcroft, P., Pau, G. S. H., Ricciuto,

- D. M., Schaefer, K., Schwalm, C. R., Serbin, S., Shevliakova, E., Slater, A. G., Tang, J., Williams, M., Xia, J., Xu, C., Joseph, R., and Koch, D. (2016). International Land Model Benchmarking (ILAMB) 2016 Workshop Report. Technical report, DOI: 10.2172/1330803.
- Hong, E., Perkins, R., and Trainor, S. (2014). Thaw settlement hazard of permafrost related to climate warming in Alaska. *Arctic*, 67(1):93–103, DOI: 10.14430/arctic4368.
- Jakeman, A. J., Letcher, R. A., and Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling and Software*, 21(5):602–614, DOI: 10.1016/j.envsoft.2006.01.004.
- Jordan, R. E. (1991). A one-dimensional temperature model for a snow cover: Technical documentation for sntherm 89. Technical report, Cold Regions Research and Engineering Laboratory, U.S. Army Office of the Chief of Engineers, United States Corps of Engineers, Hanover, NH, USA.
- Jordan, R. E., Andreas, E. L., and Makshtas, A. P. (1999). Heat budget of snow-covered sea ice at North Pole 4. *Journal of Geophysical Research: Oceans*, 104(C4):7785–7806, DOI: 10.1029/1999JC900011.
- Kobayashi, S., Ota, Y., Harada, Y., Ebata, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Kiyotoshi, T. (2015). The JRA-55 reanalysis: general specifications and basic characteristics. *Journal of the Meteorological Society of Japan*, 93(1):5–48, DOI: 10.2151/jmsj.2015-001.
- Krause, P., Boyle, D. P., and Båse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in Geosciences*, 5:89–97, DOI: 10.5194/adgeo-5-89-2005.
- Lawrence, D. M. and Slater, A. G. (2005). A projection of severe near-surface permafrost degradation during the 21st century. *Geophysical Research Letters*, 32(24):1–5, DOI: 10.1029/2005GL025080.
- Legates, D. R. and McCabe, G. J. (1999). Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1):233–241, DOI: 10.1029/1998WR900018.
- Legates, D. R. and McCabe, G. J. (2013). A refined index of model performance: A rejoinder. *International Journal of Climatology*, 33(4):1053–1056, DOI: 10.1002/joc.3487.
- Luo, D., Jin, H., Marchenko, S. S., and Romanovsky, V. E. (2018). Difference between near-surface air, land surface and ground surface temperatures and their influences on the frozen ground on the Qinghai-Tibet Plateau. *Geoderma*, 312:74–85, DOI: 10.1016/j.geoderma.2017.09.037.

- Nash, J. and Sutcliffe, J. (1970). River flow forecasting through conceptual models, part I - A discussion of principles. *Journal of Hydrology*, 10(3):282–290.
- Nazarenko, L., Schmidt, G. A., Miller, R. L., Tausnev, N., Kelley, M., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer, S., Bleck, R., Canuto, V., Cheng, Y., Clune, T. L., Del Genio, A. D., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., Legrande, A. N., Lerner, J., Lo, K. K., Menon, S., Oinas, V., Perlwitz, J., Puma, M. J., Rind, D., Romanou, A., Sato, M., Shindell, D. T., Sun, S., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M. S., and Zhang, J. (2015). Future climate change under RCP emission scenarios with GISS ModelE2. *Journal of Advances in Modeling Earth Systems*, 7(1):244–267, DOI: 10.1002/2014MS000403.
- Ochoa, V. and Urbina-Cardona, N. (2017). Tools for spatially modeling ecosystem services: Publication trends, conceptual reflections and future challenges. *Ecosystem Services*, 26:155–169, DOI: 10.1016/j.ecoser.2017.06.011.
- Oreskes, N., Shrader-Frechette, K., and Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641–646, DOI: 10.1126/science.263.5147.641.
- Osterkamp, T. E. and Romanovsky, V. E. (1999). Evidence for warming and thawing of discontinuous permafrost in Alaska. *Permafrost and Periglacial Processes*, 10(1):17–37, DOI: 10.1002/(SICI)1099-1530(199901/03)10:1<17::AID-PPP303>3.0.CO;2-4.
- Piper, M. (2018). Permamodel/Pbs V1.0. DOI: 10.5281/zenodo.1410622.
- Pomeroy, J., Gray, D., and Landine, P. (1993). The Prairie Blowing Snow Model: characteristics, validation, operation. *Journal of Hydrology*, 144(1-4):165–192, DOI: 10.1016/0022-1694(93)90171-5.
- Reichle, R. H., Draper, C. S., Liu, Q., Giroto, M., Mahanama, S. P., Koster, R. D., and De Lannoy, G. J. (2017). Assessment of MERRA-2 land surface hydrology estimates. *Journal of Climate*, 30(8):2937–2960, DOI: 10.1175/JCLI-D-16-0720.1.
- Riseborough, D., Shiklomanov, N., Etzelmüller, B., Gruber, S., and Marchenko, S. (2008). Recent advances in permafrost modelling. *Permafrost and Periglacial Processes*, 19(2):137–156, DOI: 10.1002/ppp.615.
- Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling*, 90(3):229–244, DOI: 10.1016/0304-3800(95)00152-2.
- Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., Liu, H., Stokes, D., Grumbine, R., Gayno, G., Wang, J., Hou, Y. T., Chuang, H. Y., Juang, H. M. H., Sela, J., Iredell, M., Treadon, R., Kleist, D., Van Delst, P., Keyser, D., Derber, J., Ek, M., Meng, J., Wei, H., Yang, R., Lord, S., Van Den Dool,

- H., Kumar, A., Wang, W., Long, C., Chelliah, M., Xue, Y., Huang, B., Schemm, J. K., Ebisuzaki, W., Lin, R., Xie, P., Chen, M., Zhou, S., Higgins, W., Zou, C. Z., Liu, Q., Chen, Y., Han, Y., Cucurull, L., Reynolds, R. W., Rutledge, G., and Goldberg, M. (2010). Supplement to the NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91(8):9–25, DOI: 10.1175/2010BAMS3001.2.
- Schaeffli, B. and Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, 21(15):2075–2080, DOI: 10.1002/hyp.
- Schuur, E. A., McGuire, A. D., Schädel, C., Grosse, G., Harden, J. W., Hayes, D. J., Hugelius, G., Koven, C. D., Kuhry, P., Lawrence, D. M., Natali, S. M., Olefeldt, D., Romanovsky, V. E., Schaefer, K., Turetsky, M. R., Treat, C. C., and Vonk, J. E. (2015). Climate change and the permafrost carbon feedback. *Nature*, 520(7546):171–179, DOI: 10.1038/nature14338.
- Simmons, A. J., Berrisford, P., Dee, D. P., Hersbach, H., Hirahara, S., and Thépaut, J. N. (2017). A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quarterly Journal of the Royal Meteorological Society*, 143(702):101–119, DOI: 10.1002/qj.2949.
- Smith, S. L., Romanovsky, V. E., Lewkowicz, A. G., Burn, C. R., Allard, M., Clow, G. D., Yoshikawa, K., and Throop, J. (2010). Thermal state of permafrost in North America: a contribution to the international polar year. *Permafrost and Periglacial Processes*, 21(2):117–135, DOI: 10.1002/ppp.690.
- Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M. M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M. (2013). *Climate Change 2013, the Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. DOI: 10.1017/CBO9781107415324.
- Tao, J., Reichle, R. H., Koster, R. D., Forman, B. A., and Xue, Y. (2017). Evaluation and enhancement of permafrost modeling with the NASA catchment land surface model. *Journal of Advances in Modeling Earth Systems*, 9(7):2771–2795, DOI: 10.1002/2017MS001019.
- United States Arctic Research Commission Permafrost Task Force (2003). Climate Change, Permafrost, and Impacts on Civil Infrastructure: U.S. Arctic Research Commission, Permafrost Task Force Report. Technical report, U.S. Arctic Research Commission, Arlington, Virginia.
- Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de Pison, F. J., Kaspar, F., and Sanz-Garcia, A. (2018). Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. *Solar Energy*, 164:339–354, DOI: 10.1016/j.solener.2018.02.059.

- Van Everdingen, R. (2005). *Multi-language Glossary of Permafrost and Related Ground-ice Terms (Revised)*. The Arctic Institute of North America, University of Calgary, Calgary, Alberta.
- van Vliet, J., Bregt, A. K., Brown, D. G., van Delden, H., Heckbert, S., and Verburg, P. H. (2016). A review of current calibration and validation practices in land-change modeling. *Environmental Modelling and Software*, 82:174–182, DOI: 10.1016/j.envsoft.2016.04.017.
- Van Waveren, R. H., Groot, S., Scholten, H., Van Geer, F., Wösten, H., Koeze, R., and Noort, J. (1999). *Good Modelling Practice Handbook RWS-RIZA*. Dutch Dept. of Public Works, Institute for Inland Water Management and Waste Water Treatment, 99 edition.
- Wang, W., Rinke, A., Moore, J. C., Cui, X., Ji, D., Li, Q., Zhang, N., Wang, C., Zhang, S., Lawrence, D. M., McGuire, A. D., Zhang, W., Delire, C., Koven, C., Saito, K., MacDougall, A., Burke, E., and Decharme, B. (2016). Diagnostic and model dependent uncertainty of simulated Tibetan permafrost area. *Cryosphere*, 10(1):287–306, DOI: 10.5194/tc-10-287-2016.
- Watterson, I. G. (1996). Non-dimensional measures of climate model performance. *International Journal of Climatology*, 16(4):379–391, DOI: 10.1002/(SICI)1097-0088(199604)16:4<379::AID-JOC18>3.0.CO;2-U.
- Westermann, S., Langer, M., Boike, J., Heikenfeld, M., Peter, M., Etzelmüller, B., and Krinner, G. (2016). Simulating the thermal regime and thaw processes of ice-rich permafrost ground with the land-surface model CryoGrid 3. *Geoscientific Model Development*, 9(2):523–546, DOI: 10.5194/gmd-9-523-2016.
- Westermann, S., Schuler, T. V., Gislén, K., and Etzelmüller, B. (2013). Transient thermal modeling of permafrost conditions in Southern Norway. *The Cryosphere*, 7(2):719–739, DOI: 10.5194/tc-7-719-2013.
- Williams, P. J. and Smith, M. W. (1990). *The Frozen Earth: Fundamentals of Geocryology*, volume 15. Cambridge University Press, Cambridge, DOI: 10.2307/622856.
- Willmott, C. J. (1981). On the validation of models. *Physical Geography*, 2(2):184–194, DOI: 10.1080/02723646.1981.10642213.
- Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R., O'Donnell, J., and Rowe, C. M. (1985). Statistics for the evaluation and comparison of models. *Journal of Geophysical Research*, 90(C5):8995, DOI: 10.1029/jc090ic05p08995.
- Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, DOI: 10.3354/cr030079.

- Willmott, C. J., Matsuura, K., and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, 43(3):749–752, DOI: 10.1016/j.atmosenv.2008.10.005.
- Willmott, C. J., Robeson, S. M., and Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, 32:2088–2094, DOI: 10.1002/joc.3487.
- Willmott, C. J., Robeson, S. M., Matsuura, K., and Ficklin, D. L. (2015). Assessment of three dimensionless measures of model performance. *Environmental Modelling and Software*, 73:167–174, DOI: 10.1016/j.envsoft.2015.08.012.
- Yang, K. and Zhang, J. (2018). Evaluation of reanalysis datasets against observational soil temperature data over China. *Climate Dynamics*, 50(1-2):317–337, DOI: 10.1007/s00382-017-3610-4.
- Yang, M., Nelson, F. E., Shiklomanov, N. I., Guo, D., and Wan, G. (2010). Permafrost degradation and its environmental effects on the Tibetan Plateau: A review of recent research. *Earth-Science Reviews*, 103(1-2):31–44, DOI: 10.1016/j.earscirev.2010.07.002.

Appendix A

Table of permafrost model validation procedures from nine publications

---

Surface energy transfer equation used to drive “Subsurface thermal model”

Hinzman et al. (1998)

Validation term: Validation

Data: Collected hourly measurements at seven sites of air temperature, relative humidity, wind speed, incoming shortwave radiation, and net radiation covering 4.5 months. Important meteorological variables and soil thermal properties are extrapolated to measurements on a 1km grid using a method specific to the type of measurement (air temperature accounts for elevation differences using dry or wet adiabatic lapse rate, incident shortwave radiation considers slope effect).

Procedure: Extrapolated data validated by comparison against collected data near Happy Valley site. $RMSE$, MBE , r^2 , Relative Index of Agreement (I_a) values calculated and compared for net radiation, air temperature and incoming shortwave radiation. Validation of the combined one-dimensional surface energy balance and subsurface heat transfer model was carried out by calculating $RMSE$, MBE , r^2 and I_a for measurements and simulated results near two sites using data for five different land surface types. Values used in the simulation for the thermal properties of soils were based on pre-calculated values, not measured values. Thaw depth simulated values were validated using $RMSE$, MBE , r^2 and I_a against probe measurements collected throughout the summer where five 1 km² grids were probed at 100 m increments several (3-4) times throughout the summer to determine spatial and temporal variation in depth of thaw, with the grid average at each site compared against model simulations.

Goodrich and the Guymon-Hromadka model

Osterkamp and Romanovsky (1999)

Validation term: Comparison

Data: Mean daily temperatures for 1995-97 for one depth (0.02m) at one site, mean annual observed and modelled temperature profiles for one site in 1996 compared.

Procedure: No test procedure elaborated. Two comparisons of modelled and observed temperatures presented graphically, but no exposition presented on methods of comparison besides visual.

Community Climate System Model, version 3 (CCSM3)

Lawrence and Slater (2005)

Validation term: Not discussed

Data: Qualitative against IPA, CALM ALT values for Alaskan Arctic

Procedure: Qualitative comparison to IPA map, comparison of ALT with CALM network ALT values for Alaskan Arctic

---

The Community Climate System Model, version 4 (CCSM4)

Gent et al. (2011)

Validation term: Compare

Data: Qualitative against IPA extent, CALM ALT global values. 202 CALM sites, each with between 1 and 20 years of data.

Procedure: Qualitative comparison to IPA extent map, Average of site-average ALTs for CALM sites within each grid cell compared to modelled climatological ALT for same grid cell. Borehole observations within single grid are averaged by depth and compared against corresponding model layer results.

The community climate system model version 4

Guo et al. (2012)

Validation term: Qualitatively compared, Roughly compared

Data: Data on change in permafrost temperature from 1996 to 2006 using dataset from 10 boreholes at 1 and 6m depths.

Procedure: Temperatures for 1 and 6m were estimated from model results from nearby layers through linear interpolations. The mean temperature ranges between all 10 boreholes at 1m and 6m depth calculated and compared against temperature ranges of simulated grid means. Range of simulated increases of soil temperatures at 1m and 6m were directly compared against range of observed temperature changes at those depths for 4 sites with enough data.

A new modeling approach for improved ground temperature profile determination

Badache et al. (2016)

Validation term: Comparison/Agreement

Data: Temperature from borehole in Varennes, Montreal collected from Feb-Nov 2014, confirmed with previously published data from Collins, Colorado and Temple, Texas

Procedure: Data graphed and qualitatively compared. RMSE and Normalized RMSE were calculated per depth. RMSE was 2.3K at worst, and less than 1.5K at best, NRMSE is less than 0.92%. These results are deemed satisfactory for Varennes site.

CryoGRID 1

Gisnas et al. (2016)

Validation term: Evaluation/Validation

Data: GSTs from two 500m x 500m sites with 26 and 41 data loggers respectively. MAGST calculated from temperature data from 128 loggers at all vegetation classes used in the model that cover spatially large parts of Norway. Four years of data from 25 boreholes used to evaluate modelled MAGT.

Procedure: Results of permafrost model are evaluated against average MAGST and MAGT within each grid cell, as well as the fraction of sub-zero MAGST. Lower limits of permafrost compared to permafrost probabilities derived from basal temperature of snow surveys. $RMSE$, R^2 and ME of modelled results of MAGT and MAGST against boreholes and GST loggers respectively are presented. An acceptable range of deviation is set for boreholes, percentage of borehole MAGTs falling within this range of modelled temperatures is reported as a metric.

Percentage of MAGST loggers indicating temperature below 0°C compared against percentage of area in the grid cell whose MAGST is below 0°C as generated by the model.

CryoGRID 3

Westermann et al. (2016)

Validation term: Validation

Data: Validation runs for 1979-2014 forced by measured Samoylov AWS and downscaled ERA-Interim reanalysis data.

Procedure: Validation run conducted for subsurface stratigraphy typical for Samoylov Island to compare to in situ observations. Comparisons offered are largely qualitative, conclude that satisfactory agreement of surface energy balance and surface temperatures indicates that CryoGrid 3 can consistently reproduce the land–atmosphere interactions on Samoylov Island.

NASA Catchment Land Surface Model

Tao et al. (2017)

Validation term: Validation

Data: Measurements from 51 active permafrost sites in Alaska, soil temp profiles down to 0.5–3 m. Data was screened for validity and systematic errors.

Procedure: Daily estimates of simulated soil temperature compared against observations of permafrost sites. Calculate $RMSE$ values for each layer and single, vertically averaged $RMSE$ value for each site with weights given by the layer thicknesses (deeper = thicker = more weight). Sites within same 9 km grid can exhibit large variation (0.83-6.52°C)

Table A.1: *Model result validation procedures as carried out by nine publications focusing on permafrost.*

Appendix B

Terrain type parameters

The terrain type classification scheme and parameters used in this work are those used in Cao et al. (2019). Ground types are defined by characteristics chosen for their strong connection to observed spatial variation of GST, as well as the ease of their identification in the field. Figure B.1 from Cao et al. (2019) is a visual representation of each of the ten ground types.

Figure B.1: The 10 ground types used to group observations, modified from Cao et al. (2019) to use the term *Till Veneer* instead of *Till Blanket* for ground type (e).

Soil stratigraphy parameters of each ground type are defined to match one of five soil profiles, seen in Figure B.2 from Cao et al. (2019), each of which is a different configuration of three of the six subsurface materials whose GEOTop parameters are defined in B.1 from Cao et al. (2019). These soil parameters are used to define the properties of the soil layers

Table B.1: GEOTop soil parameters for different subsurface materials from Cao et al. (2019).

Stratigraphy	Glaciofluvial	Sandy till	Silty till	Peat	Ice-wedge	Bedrock
Saturated water content (%)	37.5	50	39.7	85	95	5
Residual water content (%)	4.9	4.9	5.3	20	0	1
Van Genuchten α (m^{-1})	2.6	2.1	0.8	30	30	1
Van Genuchten $n(-)$	1.79	1.57	1.51	1.8	1.8	1.2
Thermal capacity ($10^6 \text{J m}^{-3} \text{K}^{-1}$)	2	2	2	1.8	1.8	2
Thermal conductivity ($\text{W m}^{-1} \text{K}^{-1}$)	4	2	2	0.22	0.22	2

being used in the simulation, and along with the snow and vegetation parameters listed in Table B.2 are included in the GEOTop configuration file.

Figure B.2: Soil profiles of different ground types from Cao et al. (2019). In this simulation, the layer of bedrock is extended to -50 m. Parameters for subsurface materials are provided in Table B.1 from Cao et al. (2019).

Table B.2: *GEOtop soil, snow and vegetation parameters for different ground types from Cao et al. (2019).*

Terrain	E	SV	IWP	SG	TV	IWT	SH	SF	SD	TS	
Soil parameters											
Stratigraphy	I	II	III	III	II	IV	V	III	II	II	
Soil water content (%)	12	45	85	77.5	40	95	36	85	45	45	
	15	50	39.7	39.7	50	39.7	39.7	39.7	50	50	
	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	
Simulation depth (m)						50					
Initial ground temperature (°C)						-2					
Snow parameters											
Snow correction factor (–)	0.3	0.65	1	1.1	0.8	1.1	1.3	1.3	1.62	1.35	
Wind compaction (–)	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	
Vegetation parameters											
Vegetation height (m)	0.15	0.20	0.20	0.15	0.30	0.20	0.30	0.35	0.15	2.00	
LSAI (m ² m ^{−2})	1.5	2.3	4.1	3.7	2.7	2.1	3.5	2.1	3.0	3.8	
Canopy fraction (–)	0.4	1.0	0.8	0.8	0.7	0.8	0.8	0.6	0.8	1.0	
DecayCoeffCanopy (–)	2.5	2.7	2.7	2.5	3	2.7	2.7	3.2	2.5	3.5	
Vegetation root depth (m)	0.10	0.10	0.15	0.10	0.20	0.15	0.20	0.10	0.10	0.7	
Vegetation reflectivity in the visible (–)						0.11					
Vegetation transmissivity in the visible (–)						0.55					
Vegetation reflectivity in the near infrared (–)	0.45	0.50	0.45	0.40	0.45	0.40	0.50	0.40	0.45	0.45	
Vegetation transmissivity in the near infrared (–)	0.25	0.25	0.30	0.34	0.25	0.30	0.25	0.34	0.25	0.25	
Surface density of canopy (kg m ^{−2})	0.4	1.6	3.6	3.5	1.7	1.6	4.1	2.8	3.5	4.0	
VegSnowBurying (–)	2.0	2.0	1.5	1.0	2.0	1.5	2.0	1.0	2.0	2	

DecayCoeffCanopy is the decay coefficient of the eddy diffusivity profile in the canopy, and VegSnowBurying is the coefficient of the exponential snow burying of vegetation.

Appendix C

Annual data tables and histograms

Table C.1: Annual accordance means, ranks, and 95% confidence intervals for six models calculated over the entire dataset using *RMSE*.

Model	Mean	Rank	Confidence Interval
GT-ERA5	2.8081	1	(2.6950, 2.9242)
GT-JRA55	2.9088	3	(2.7892, 3.0397)
GT-BSnow-ERA5	4.0699	4	(3.8817, 4.2628)
GT-BSnow-JRA55	2.8237	2	(2.7077, 2.948)
ERA5/2	5.4148	6	(5.2616, 5.5734)
JRA55/2	5.2461	5	(5.1171, 5.3862)

Table C.2: Annual accordance means, ranks, and 95% confidence intervals for six models calculated over the entire dataset using d_1 .

Model	Mean	Rank	Confidence Interval
GT-ERA5	0.8173	1	(0.8085, 0.8259)
GT-JRA55	0.8039	3	(0.7947, 0.8125)
GT-BSnow-ERA5	0.7683	4	(0.7565, 0.7799)
GT-BSnow-JRA55	0.8138	2	(0.8048, 0.8222)
ERA5/2	0.6718	6	(0.6589, 0.6839)
JRA55/2	0.6723	5	(0.6608, 0.6832)

Table C.3: Annual accordance means, ranks, and 95% confidence intervals for six models calculated over the entire dataset using E_1 .

Model	Mean	Rank	Confidence Interval
GT-ERA5	0.5927	2	(0.5664, 0.6159)
GT-JRA55	0.5919	3	(0.5694, 0.6122)
GT-BSnow-ERA5	0.3909	4	(0.3499, 0.4291)
GT-BSnow-JRA55	0.6013	1	(0.5778, 0.6226)
ERA5/2	0.1695	6	(0.1089, 0.2207)
JRA55/2	0.1931	5	(0.1385, 0.2404)

Figure C.1: Histograms of r^2 over six models using annual datasets. The dashed red line represents the per-model mean of the r^2 values over all point-years in the dataset.

Figure C.2: Histograms of MAE over six models using annual datasets. The dashed red line represents the per-model mean of the MAE values over all point-years in the dataset.

Figure C.3: Histograms of RMSE over six models using annual datasets. The dashed red line represents the per-model mean of the RMSE values over all point-years in the dataset.

Figure C.4: Histograms of NSE over six models using annual datasets. The red dashed line represents the per-model mean of the NSE values over all point-years in the dataset.

Figure C.5: Histograms of d over six models using annual datasets. The red dashed line represents the per-model mean of the d values over all point-years in the dataset.

Figure C.6: Histograms of d_1 over six models using annual datasets. The red dashed line represents the per-model mean of the d_1 values over all point-years in the dataset.

Figure C.7: Histograms of d_r over six models using annual datasets. The red dashed line represents the per-model mean of the d_r values over all point-years in the dataset.

Figure C.8: Histograms of E_1 over six models using annual datasets. The red dashed line represents the per-model mean of the E_1 values over all point-years in the dataset.

Appendix D

Seasonal data accordances and ranks

Table D.1: Mean accordance of all models over annual dataset compared with mean accordances of all models over seasonal datasets.

Accordance Measure	Annual	Seasonal Mean	Summer	Fall	Winter	Spring
MAE	3.0516	3.0634	2.8833	1.9404	4.4906	2.9393
$RMSE$	3.8786	3.5351	3.2425	2.4255	4.9270	3.5454
NSE	0.4668	-63.3295	-1.2318	-0.9650	-244.3054	-6.8156
d	0.9124	0.743375	0.7420	0.8282	0.5827	0.8206
d_1	0.7579	0.55855	0.5317	0.6505	0.4069	0.6451
d_r	0.7120	0.35205	0.3576	0.4758	0.0540	0.5208
E_1	0.4232	-1.23133	-0.3467	-0.0925	-4.1710	-0.3151

Appendix E

Spatial data accordances and ranks

Figure E.1: Plausible model ranks for each model in summer based on accordance measures RMSE, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), and Snowdrifts (SD). *cont'd*

Figure E.1: Plausible model ranks for each model in summer based on accordance measures RMSE, d_1 and E_1 on ground types Sedge fen (SF), Sedges and grass (SG), Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).

Figure E.2: Plausible model ranks for each model in fall based on accordance measures RMSE, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), Snowdrifts (SD), Sedge fen (SF), and Sedges and grass (SG). *cont'd*

Figure E.2: Plausible model ranks for each model in fall based on accordance measures $RMSE$, d_1 and E_1 on ground types Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).

Figure E.3: Plausible model ranks for each model in winter based on accordance measures RMSE, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), Snowdrifts (SD), Sedges fen (SF), and Sedges and grass (SG). *cont'd*

Figure E.3: Plausible model ranks for each model in winter based on accordance measures $RMSE$, d_1 and E_1 on ground types *Shrubby hummocks (SH)*, *Short vegetation (SV)*, *Tall shrubs (TS)*, and *Till veneer (TV)*.

Figure E.4: Plausible model ranks for each model in spring based on accordance measures RMSE, d_1 and E_1 on ground types Esker (E), Ice-wedge polygons (IWP), Ice-wedge troughs (IWT), Snowdrifts (SD), Sedge fen (SF), and Sedges and grass (SG). *cont'd*

Figure E.4: Plausible model ranks for each model in spring based on accordance measures RMSE, d_1 and E_1 on ground types Shrubby hummocks (SH), Short vegetation (SV), Tall shrubs (TS), and Till veneer (TV).