A Statistical Ranking Framework For Ground Temperature Models, Tailored Towards Permafrost Environments.

Hannah MacDonell, Stephan Gruber
M.Sc. D.S. Carleton University

Northwest Territories | PermafrostNet NSERC | CRSNG | Digital Research Alliance of Canada

COLLABORATORS / ACKNOWLEDGEMENTS
GTPEM Development / Technical Support: Brown, N., Cao, B.
Committee Members: Gruber, S., Kokelj, S., Lantz, T., Melton, J.

12th International Conference on Permafrost 2024 XII Whitehorse Yukon

Introduction

Permafrost modelling can contribute to **informing adaptation** in permafrost regions by characterizing the current and future state of the ground. Until recently, the advancement of permafrost modelling has been **limited by sparse data for both driving models** (surface forcing meteorology) and **evaluating model predictions** (observations of the simulated variable). The emergence of reanalysis data products and enhanced data sharing has extended modelling to new permafrost regions. With this, our capacity to assess modelling applications should also improve, but unfortunately, there exist **few systematic approaches for doing so.** This study proposes a ranking framework to address challenges in evaluating models and serves as an intermediate step towards **standardizing the interpretation and comparison** of model performance using large observational datasets.

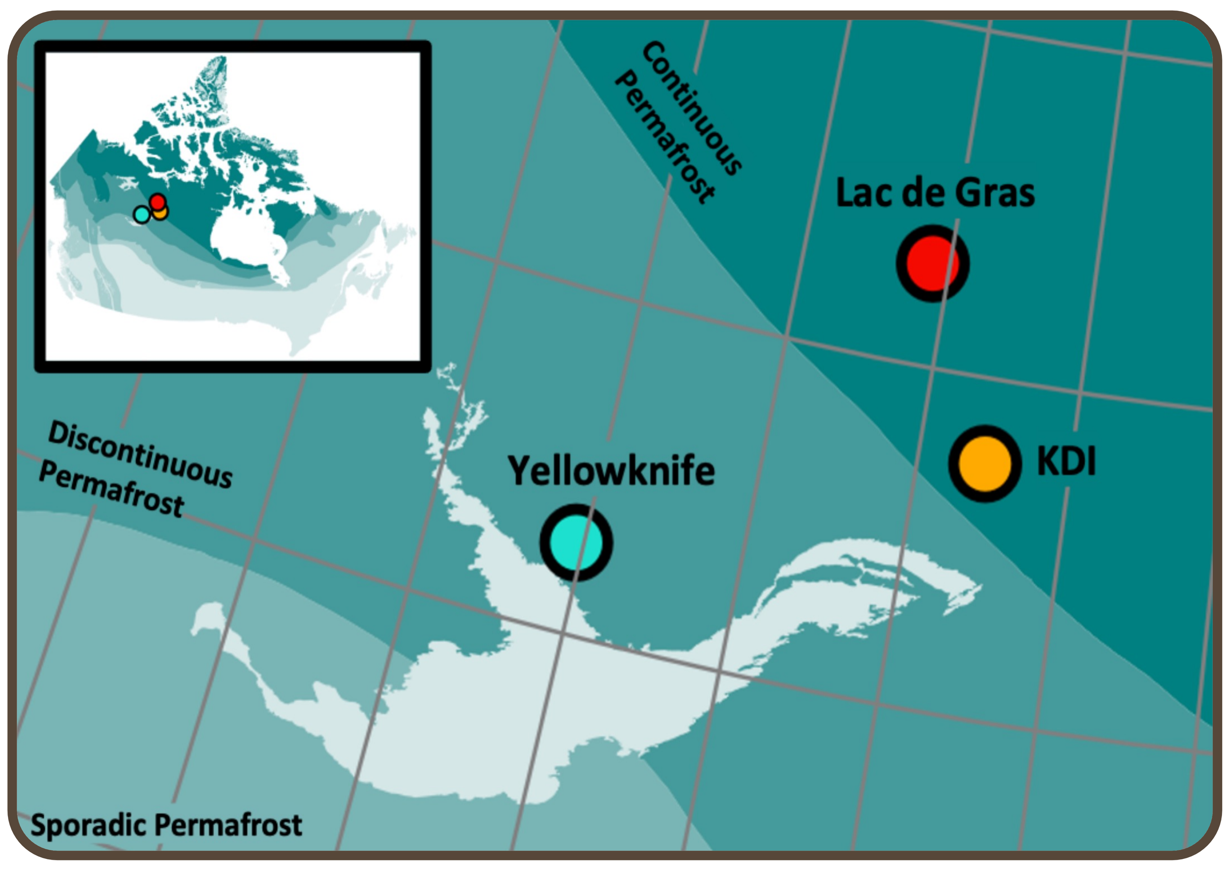


Figure 1: Location of three clusters of ground-surface temperature plots in the Northwest Territories, Canada. Temperature observations at these locations were used to evaluate model output.

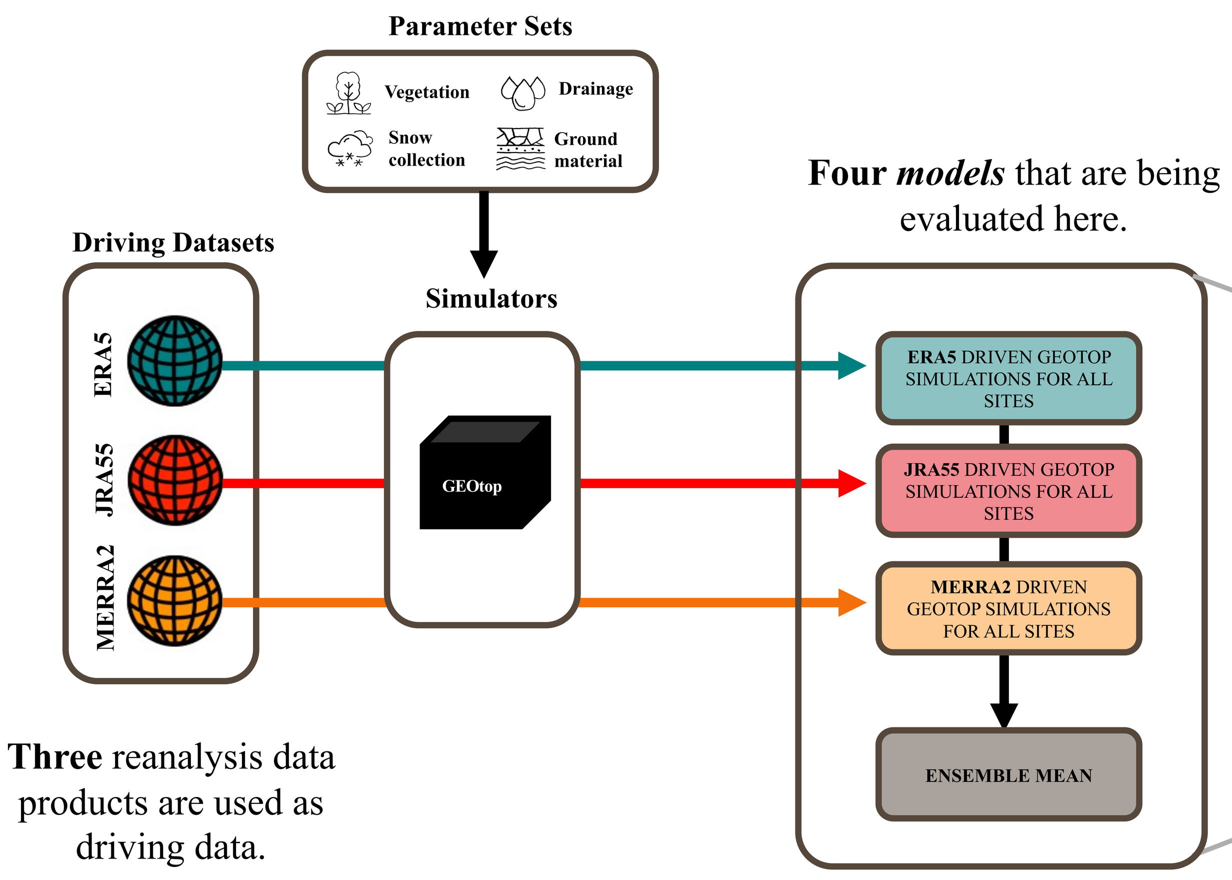
Methodology

Simulations — We are comparing different **models** which are a combination of: **driving data**, **modelling software** (GEOtop), and **parameters** (see schematic below**)**. The four different models evaluated here are represented using the symbology: $\boldsymbol{M_1}$, $\boldsymbol{M_2}$, $\boldsymbol{M_3}$, and their ensemble, $\boldsymbol{M_E}$.

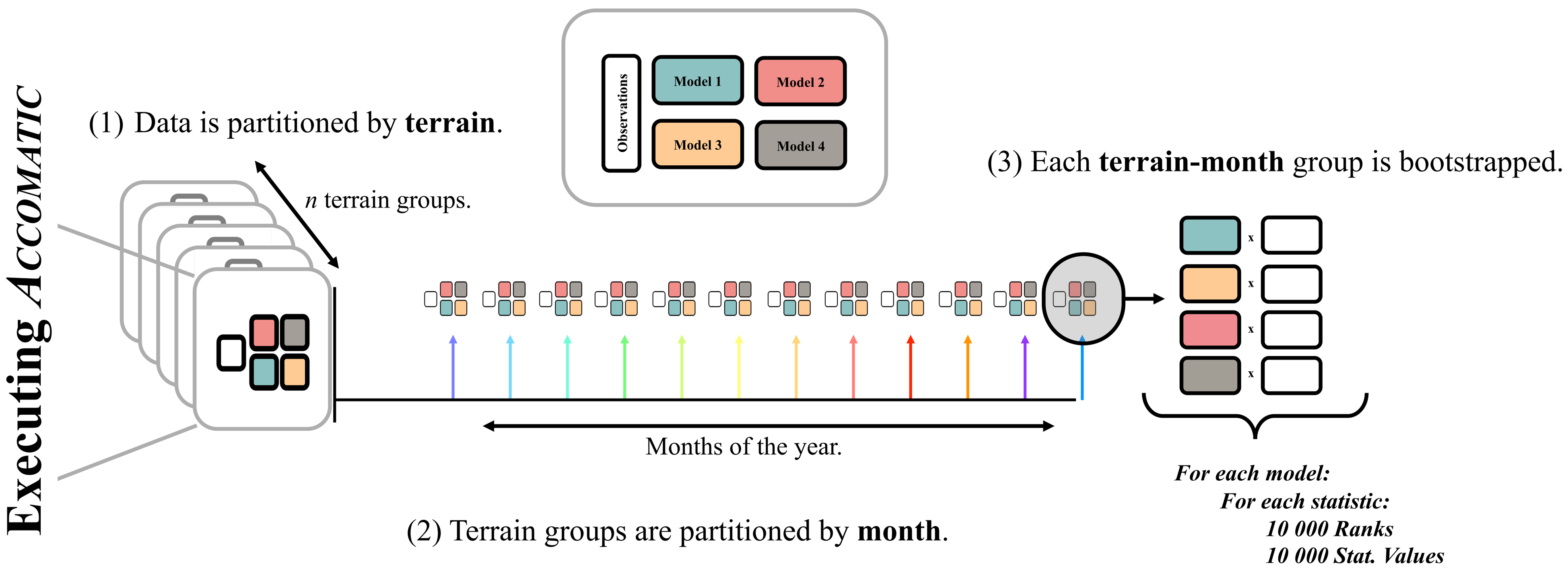
Accomatic — The python package *Accomatic* produces a suite of summary statistics and model rankings. Each model was tested with a range of accordance measures, stratified by season and terrain type.

Approach — Each **model evaluation challenge** addressed by this ranking framework is summarized below. Each of the solutions described is programmed into the model ranking tool.

Producing GST Models



Executing *Accomatic*

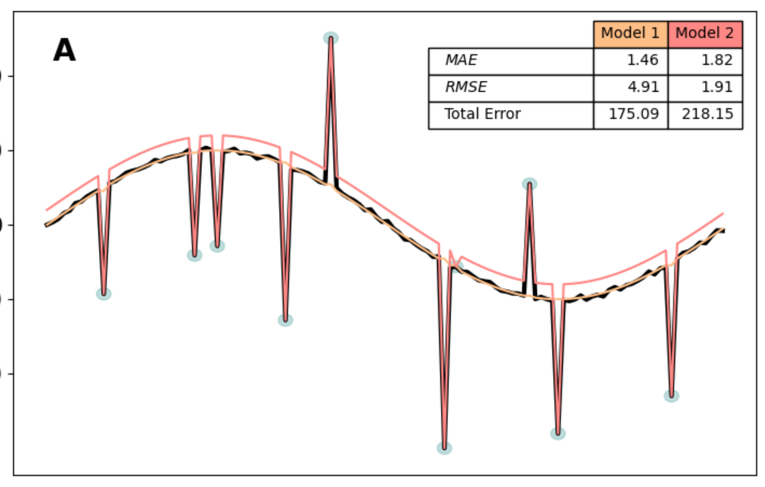


Permafrost Model Evaluation Challenges and *Accomatic* Solutions

Lack of Statistical Consensus

Problem: Models are difficult to compare due to the **lack of consensus** over which statistics to use. There are limitations to many commonly used statistics, including **A)** how large, low frequency errors are penalized, **B)** how error close to zero influences statistical results and **C)** the information they provide.

Solution: Three statistics are selected to evaluate temperature simulations: ***BIAS***, ***MAE***, and ***R***.



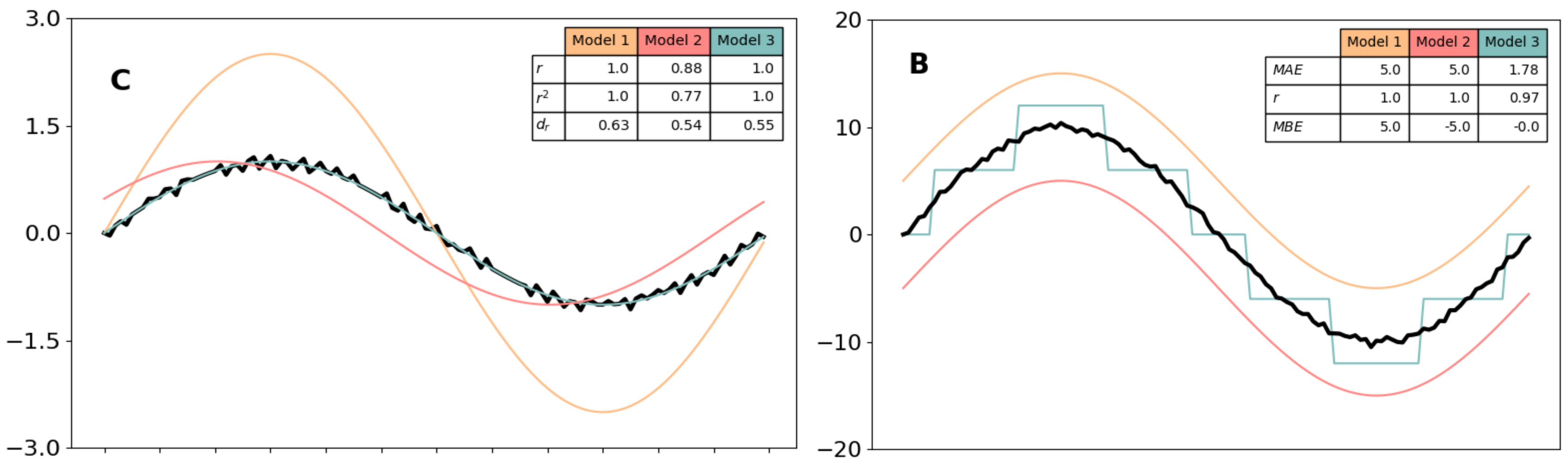
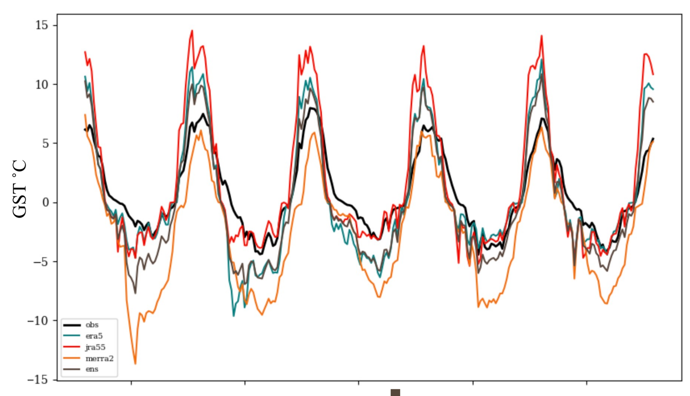


Figure 2: Demonstrating how statistical measure selection influences our interpretation of performance.

Incomplete Observational Datasets

Problem: To avoid introducing seasonal bias into model results, **complete years of data** are favoured for evaluation. This means **lots of data is lost** from model evaluation.

Solution: The bootstrap procedure implemented by **Accomatic** segments modelled and observed timeseries into month-long sections, then evaluates **random samples** from this set, getting a distribution of model performance.



A **mean** and **spread** of model performance from sampling complete months with replacement.

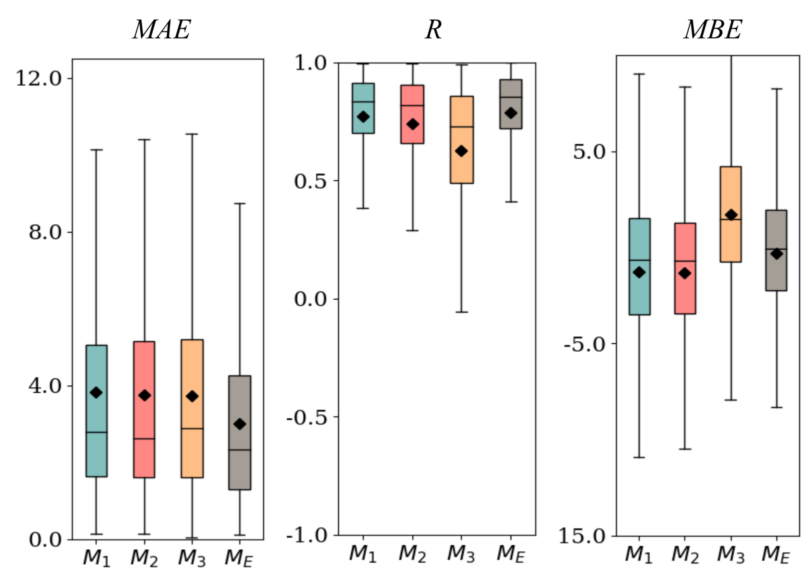


Figure 3: A schematic showing how timeseries observations and model output is summarized into boxplots showing a distribution of model performance across three different statistics (*MAE*, *R*, and *MBE*).

Limited Spatial Coverage

Problem: Permafrost environments exhibit remarkable heterogeneity and model evaluation can be biased towards areas for which we have more data.

Solution: Model evaluation is **subset** by **terrain type.** This allows for a better understanding of how the model performs in different environments, **mitigating potential bias** towards terrains with abundant observations

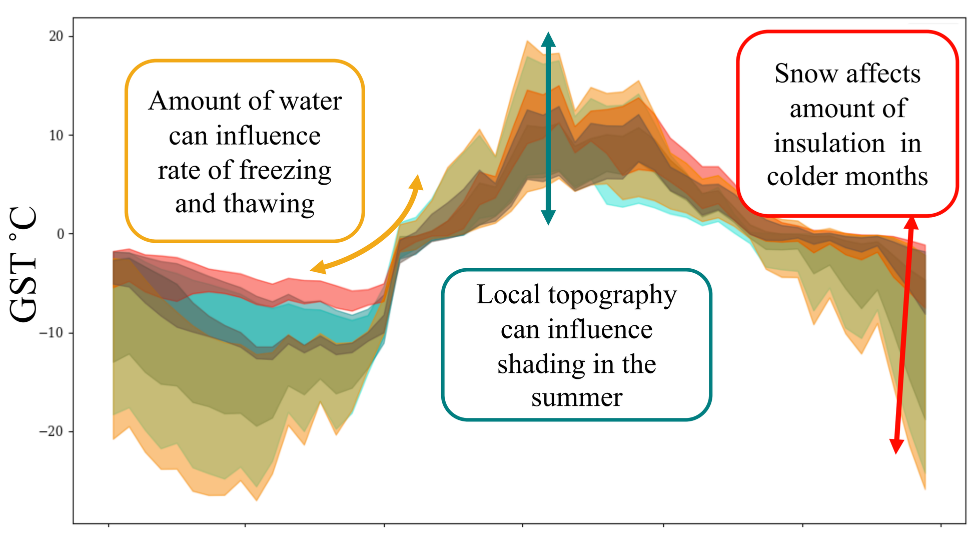


Figure 4: Visualization of how observed GST can vary across different classes of terrain.

Interpretation of Statistical Values

Problem: Most statistical values are **intangible** and often **mathematically unrelated** to one another, making them difficult to interpret.

Solution: Relative performance between models is recorded as a rank for each bootstrap sample evaluated for each terrain type and month of the year. This means thousands of ranks can be aggregates across multiple levels of evaluation to achieve a distribution of model rankings, shown in Figure 5. e.g. $\mathbf{M_E}$ ranks first most often (34%) while $\mathbf{M_3}$ ranks poorest, placing last 58% of the time. *MBE* shows the *proportion of instances* a model demonstrated warm bias.

Aggregate Rank Distribution / ***MBE***

	First	Second	Third	Fourth	*MBE*
M_1	0.27	0.22	0.31	0.2	0.42
M_2	0.16	0.25	0.37	0.23	0.4
M_3	0.23	0.08	0.11	0.58	0.66
M_E	0.34	0.45	0.21	0	0.48

Figure 5: Ranking distribution of four models across, aggregated across all months of the year and terrain types.

Future Work

- Next, this method could be **tailored to other variables** of interest as Accomatic is currently specific to ground surface temperature.
- Applying this method **using different permafrost models** (e.g. *CLASSIC*, *FreeThaw1D*)
- Incorporating this method seamlessly into a comprehensive simulation workflow.

Contact Info

Hannah Macdonell, M.Sc. D.S. Candidate. Geography Department, Carleton University
Email: Hannah.Macdonell@CARLETON.ca

Dr. Stephan Gruber, Geography Department, Carleton University
Email: StephanGruber@CUNET.CARLETON.CA