A MACHINE LEARNING SURROGATE FOR ALPINE3D SNOWPACK SIMULATION: PERSISTENCE-BASED EVALUATION AND A PATH TOWARD SNOWDRIFT

Michael Brandon Hurd[1]*

[1]*AutoRoboto, San Francisco, CA, USA*

ABSTRACT: Distributed snow-cover models such as Alpine3D and SNOWPACK provide spatially detailed information on snowpack evolution, surface energy balance, and terrain-driven variability that is relevant to avalanche operations. The drawback is the computational cost of high-resolution, multi-season simulations, which limits repeated scenario testing, ensemble analysis, and rapid visualization. This paper presents a physics-supervised graph neural network surrogate for selected gridded Alpine3D/SNOWPACK outputs over a Kananaskis Country case-study domain in the Canadian Rockies. The foundation model was trained on multiple winter seasons and evaluated on a fully held-out season. The model represents the distributed domain as a graph and uses terrain, previous snow state, meteorological forcing, and time to predict selected snow-state and process variables.

The first evaluation compared absolute snow height (HS) and snow water equivalent (SWE) predictions with a copy-previous-hour persistence baseline. Although the model achieved R^2 values of 0.99886 for HS and 0.99939 for SWE, persistence performed better, with R^2 values of 0.99992 and 0.99998. For this reason, hourly changes were reconstructed from predicted process variables. The reconstructed changes achieved positive skill against persistence of 0.373 for ΔHS and 0.795 for ΔSWE, with the strongest performance during melt and runoff. A separate frozen-checkpoint test on 290 corrected-2022 timesteps and approximately 312 million valid cell observations retained positive flux-reconstructed skill of 0.479 for ΔHS and 0.461 for ΔSWE. This provides evidence of robustness to the corrected forcing lineage within the same domain.

End-to-end inference required 23.8 s per predicted hour on the 1.12-million-cell domain, compared with 432 s per simulated hour for the measured Alpine3D workflow. This represents an approximately 18× system-level wall-clock speedup across different GPU and CPU hardware. The foundation teacher uses simple radiation and does not include wind-driven snow transport. Higher-fidelity physics extensions are being evaluated, but they remain preliminary and are not reported as validated results.

The objective is not to replace Alpine3D, SNOWPACK, or field observations. The objective is to determine which teacher-model processes can be reproduced by a graph surrogate, how that reproduction should be evaluated, and what validation is required before operational use.

KEYWORDS: graph neural network, Alpine3D, SNOWPACK, snowpack modelling, avalanche forecasting, persistence baseline

1. INTRODUCTION

Distributed snow-cover models are used to estimate spatial variation in snow accumulation, settlement, melt, runoff, and surface energy exchange. SNOWPACK resolves one-dimensional snow-cover processes, while Alpine3D applies snow and surface-process calculations across gridded mountain terrain (Bartelt and Lehning, 2002; Lehning et al., 2006). These models can support research and the operational interpretation of terrain-dependent snow conditions. The drawback is the computational cost of high-resolution, multi-season simulations. A learned surrogate could reduce this cost for repeated scenario testing, ensembles, and rapid visualization.

A gridded mountain domain can be represented as a graph, where the grid cells are nodes connected by spatial relationships. Graph-network simulators and MeshGraphNets have shown that message-passing neural networks can learn the evolution of complex physical and mesh-based systems (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021). Physics-constrained neural differential equations have also been applied to daily, site-based snow-depth simulation (Charbonneau et al., 2025). This work is different because it predicts distributed grid-cell teacher outputs and uses an hourly evaluation. At an hourly time step, snow height and snow water equivalent often change very little. Because of this, a model can achieve an apparently excellent absolute-state coefficient of determination by copying the previous state. The model may still fail to learn useful accumulation, settlement, melt, or runoff dynamics.

* Corresponding author address:
Brandon Hurd
AutoRoboto, San Francisco, CA, USA
email: brandon@autoroboto.io

The goal of this work is to determine if a graph neural network can reproduce selected outputs from distributed Alpine3D/SNOWPACK simulations. The model is evaluated by comparing absolute-state metrics with a persistence-referenced change evaluation. Changes in HS and SWE are also reconstructed from predicted teacher fluxes. The runtime is compared with production Alpine3D timing on the same full domain. Lastly, a paired-simulation design is described for adding complex-terrain radiation and wind-transport supervision. This last component is a research path and is not a validated extension.

The contribution of this work is the model and the evaluation method used to separate useful hourly prediction skill from persistence. The purpose is not to replace Alpine3D, SNOWPACK, or field observations. The purpose is to determine which parts of the physics teacher can be reproduced by the graph surrogate and what still needs to be tested before the model can be considered for operational use.

2. DATA AND METHODS

2.1 Teacher simulations and dataset

The teacher data were generated with Alpine3D/SNOWPACK over a Kananaskis Country case-study domain in the Canadian Rockies. MeteoIO was used to preprocess and spatially interpolate the meteorological forcing (Bavay and Egger, 2014). Eight winter-season datasets spanning 2017–2024 were prepared for the foundation model. Seven seasons were used for training, and the entire 2023 season was reserved for evaluation. Four of the season datasets used fitted-lapse or gap-stitched forcing reconstructions. The 2023 season was not included in the normalization statistics.

The foundation teacher used the SIMPLE terrain-radiation configuration and did not include wind-driven snow transport. This limits what the foundation surrogate can learn. The surrogate can only reproduce the processes and approximations that are present in the teacher outputs. The model is evaluated against teacher outputs from the same simulation lineage. For this reason, the results measure teacher reproduction and not accuracy against independent physical observations.

The gridded data store contains static terrain attributes, forcing variables, teacher outputs, and masks. Model samples are drawn as fixed spatial graph crops. The 2023 evaluation season contains 8,039 hourly samples. A fixed evaluation panel was created from 64 selected timesteps and approximately 4.05 million valid cell observations.

2.2 Graph neural surrogate

The surrogate uses a multi-scale message-passing graph network. Spatial connections allow information to move through local and wider terrain neighborhoods. The architecture contains encoding, graph-processing, and output-decoding components.

The model inputs describe terrain and surface conditions, the previous snow state, meteorological forcing, and temporal context. Wind speed and direction are not inputs to the foundation model. The selected teacher outputs represent snow-state and process-variable groups. The previous state and current forcing are used to make a one-step prediction. Snow-profile outputs are not evaluated in this paper.

Table 1 summarizes the model and evaluation configuration.

Table 1: Foundation-model and evaluation specification.

Item	Specification
Architecture	Multi-scale message-passing graph network with encoding, processing, and decoding components
Graph sample	Fixed spatial graph crops with local and wider terrain connections
Model contract	Terrain, prior snow state, meteorological forcing, and time inputs; selected snow-state and process outputs
Foundation teacher	SIMPLE terrain radiation; no wind input; no snow transport
Evaluation	Entire 2023 season held out and excluded from normalization; 64 timesteps and approximately 4.05 million valid cells

2.3 Training and holdout design

The model was trained with a multi-objective loss over the selected teacher variables. Training and internal validation samples were separated temporally, and a frozen checkpoint was selected before final evaluation. The 2023 holdout uses the same physical domain as the training seasons but contains no training timestamps. The holdout tests temporal generalization to a different season. It does not test geographic generalization.

All reported results are one-step evaluations initialized from the teacher state. A long autoregressive rollout of the frozen checkpoint was not tested. In this type of test, a model prediction would be used as an input for the next prediction. Since this test was not performed, the results do not establish multi-day forecast stability.

After the primary evaluation, the frozen foundation checkpoint was tested against corrected 2022 TCH forcing and teacher outputs. The evaluation used 290 stratified timesteps and approximately 312 million valid cell-time observations. The corrected-2022 samples were not used to update the checkpoint or select model parameters. Since the 2022 season had been represented in the original training corpus under the legacy forcing lineage, this evaluation tests robustness to corrected forcing within the same domain. It is not treated as an independent temporal or geographic holdout.

2.4 Persistence-referenced evaluation

For an hourly change quantity Δ, the persistence forecast is Δ = 0. Following the mean-square-error skill-score framework of Murphy (1988), skill is defined as

$$skill = 1 - \frac{MSE(model)}{MSE(persistence)} \qquad (1)$$

A value greater than zero improves on persistence; zero is equivalent to persistence; and a negative value is worse.

The first evaluation used R^2 to compare the absolute HS and SWE predictions with a copy-previous-hour baseline. The next step was to evaluate the hourly changes in two ways. The direct-state change is the predicted absolute state minus the previous state. The flux-reconstructed change combines selected mass and height-change outputs using the teacher model's units and sign conventions. The normalized predictions are returned to physical units before the changes are reconstructed. On the fixed panel, the true teacher fluxes close against the teacher-state changes with correlations of 0.9966 for HS and 0.9910 for SWE. The magnitude ratios show that the hourly output terms do not provide exact budget closure.

Separate metrics were calculated for accumulation, melt, runoff, snowline transition, and elevation subsets. A composite evaluation was used to summarize performance across the principal state-change and regime subsets.

2.5 Staged radiation and transport supervision

The foundation teacher does not include complex terrain radiation or wind-driven redistribution. Complex terrain changes shortwave radiation through shading, restricted sky view, and terrain reflection (Helbig and Löwe, 2012). Saltation, suspension, and preferential deposition also contribute to mountain snow redistribution (Mott et al., 2010). A controlled paired-simulation workflow was established at higher spatial resolution to evaluate these omitted processes while holding the underlying forcing and baseline configuration constant. One paired season was completed, but later corrections to simulation outputs and unavailable verification records prevent this dataset from supporting a validated model result. The completed workflow is reported only as preliminary future work.

2.6 Runtime benchmarking

Runtime was tested on the full 1113 × 1004 Kananaskis domain, which contains 1,117,452 cells at 100 m resolution. The GNN benchmark used an NVIDIA RTX A6000 and single-precision inference for 12 predicted hours. Two GNN times are reported. The model-only time includes the network execution, while the end-to-end time also includes data loading, graph construction, data transfer, and output assembly. The Alpine3D timing was reconstructed from 1,920 production hourly steps on an Intel Xeon w9-3475X. Additional tests across several graph sizes were used to characterize latency scaling.

The runtime test is a system-level wall-clock comparison and does not compare identical numerical work. The GNN predicts a selected subset of teacher outputs, while Alpine3D advances the full configured physical model. The two systems also use different processor types. The foundation comparison uses SIMPLE radiation without snow transport. For this reason, the result measures the speed of the current workflows on a shared domain and forcing lineage. It is not a functionally equivalent or hardware-normalized comparison.

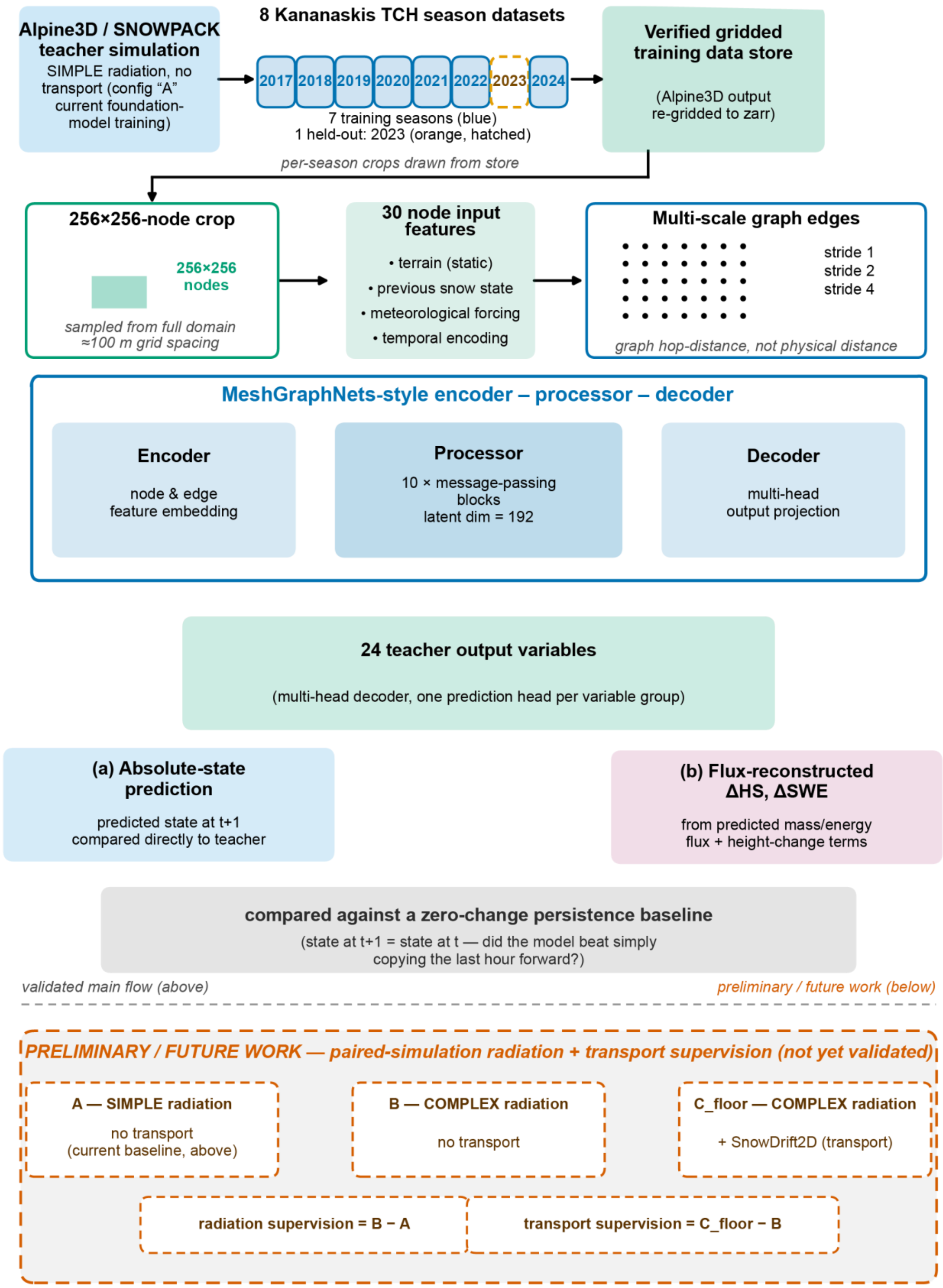

Figure 1: Architecture and evaluation workflow. The foundation model is trained on simple-radiation, no-transport teacher simulations and evaluated through both absolute-state and persistence-referenced change metrics. Higher-fidelity physics extensions remain preliminary and are not included in the validated results.

3. RESULTS

3.1 *Absolute-state and persistence testing*

The first test compares the absolute-state predictions with persistence. The frozen foundation checkpoint was evaluated on the fixed 2023 holdout panel, which contains 4,050,560 valid cell observations. The absolute-state R^2 is 0.99886 for HS and 0.99939 for SWE. These values appear very high, but copying the previous hour produces higher values of 0.99992 for HS and 0.99998 for SWE (Table 2).

The problem is easier to see when the predicted absolute states are converted to hourly changes. The direct state-implied skill relative to persistence is −13.70 for ΔHS and −36.19 for ΔSWE. The absolute-state head does not provide useful hourly change estimates even though the R^2 is nearly one. The reason for this behavior is that HS and SWE change slowly over one hour and their previous values are available as model inputs. The result is not caused by timestamp overlap with the held-out season.

3.2 *Flux-reconstructed change testing*

The next test reconstructs the hourly changes from predicted process variables. This produces a different result. Skill relative to persistence is 0.373 for ΔHS and 0.795 for ΔSWE, with correlations of 0.607 and 0.907. The positive skill shows that the model learned useful teacher tendencies. The absolute-state decoder does not express these tendencies reliably when consecutive absolute-state predictions are subtracted.

The performance changes with the snow regime. Melt-regime SWE skill is 0.961, and runoff skill is 0.936. Accumulation skill is 0.411 for HS and 0.513 for SWE. Snowline skill is 0.703. The weakest subset is low-elevation HS, with a skill of 0.215 compared with 0.454 at high elevations. From these results, thin or marginal snow cover near the rain–snow transition is harder to predict than deeper alpine snow in the sampled domain.

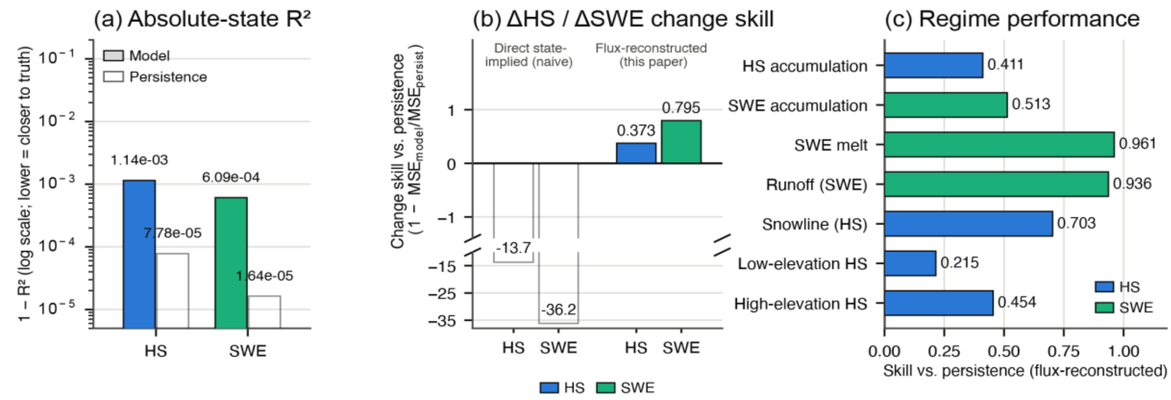

Figure 2: Persistence-corrected evaluation on the fixed 2023 holdout panel. Absolute-state R^2 is high but below a copy-previous-hour baseline, while flux-reconstructed changes achieve positive skill for both HS and SWE.

Table 2: Fixed-panel performance on the held-out 2023 season. Skill is measured against a zero-change persistence forecast.

Metric	HS	SWE
Model absolute-state R^2	0.99886	0.99939
Persistence absolute-state R^2	0.99992	0.99998
Direct state-implied change skill	−13.70	−36.19
Flux-reconstructed change skill	0.373	0.795
Flux-reconstructed change correlation	0.607	0.907
Teacher-flux closure correlation	0.9966	0.9910
Accumulation-regime skill	0.411	0.513

The results show that absolute-state accuracy is not an adequate primary metric for this hourly surrogate. The decomposed teacher targets retain useful dynamical information even when persistence dominates the absolute-state decoder.

3.3 *Corrected-2022 robustness test*

The frozen foundation checkpoint was also evaluated on the corrected 2022 TCH dataset. The evaluation contained 290 stratified timesteps and approximately 312 million valid cell-time observations. The canonical flux reconstruction achieved R^2 values of 0.475 for ΔHS and 0.451 for ΔSWE. Skill relative to persistence was 0.479 for ΔHS and 0.461 for ΔSWE.

The absolute-state evaluation showed the same persistence-inflation behavior observed in the 2023 holdout. The absolute HS prediction achieved an R^2 of 0.9976, but persistence remained stronger. This result provides an additional check that a high absolute-state R^2 does not establish useful hourly prediction skill.

Performance varied by regime. The strongest results occurred during storm-related conditions. Skill was 0.738 for ΔHS and 0.813 for ΔSWE during storm-to-melt transitions, and 0.587 for ΔHS and 0.707 for ΔSWE during early-season storms. Baseline and other non-storm regimes retained positive but more modest skill. Cold, clear midwinter conditions were the main exception, with skill of −1.88 for ΔHS and −6.03 for ΔSWE over approximately 5.4 million valid cell observations.

The corrected-2022 result shows that the process tendencies remain useful after the forcing correction. Since the same season was represented in the legacy training lineage, the result is evidence of robustness to corrected forcing rather than independent temporal transfer.

Table 3: Corrected-2022 performance of the frozen foundation checkpoint. Skill is measured against a zero-change persistence forecast.

Metric or regime	HS	SWE
Flux-reconstructed change R^2	0.475	0.451
Flux-reconstructed change skill	0.479	0.461
Storm-to-melt transition skill	0.738	0.813
Early-season storm skill	0.587	0.707
Cold, clear midwinter skill	−1.88	−6.03

3.4 Full-domain runtime

The mean model-only execution time for the full domain was 0.885 ± 0.026 s per predicted hour (n = 12). The end-to-end execution time was 23.83 ± 1.72 s per predicted hour. Data loading and device-to-host output transfer account for most of the non-model time. The corresponding Alpine3D production mean was 432.5 s per simulated hour, with a median of 358.3 s over 1,920 steps. These measurements give a 489× model-only speedup and an 18.1× end-to-end system-level wall-clock speedup. The GNN latency increased approximately linearly with node count over the tested graph sizes. Since the hardware and functional scope are different, the 18.1× end-to-end value is used as the operational comparison. The 489× model-only value shows how much time could be recovered by improving the input and output pipeline.

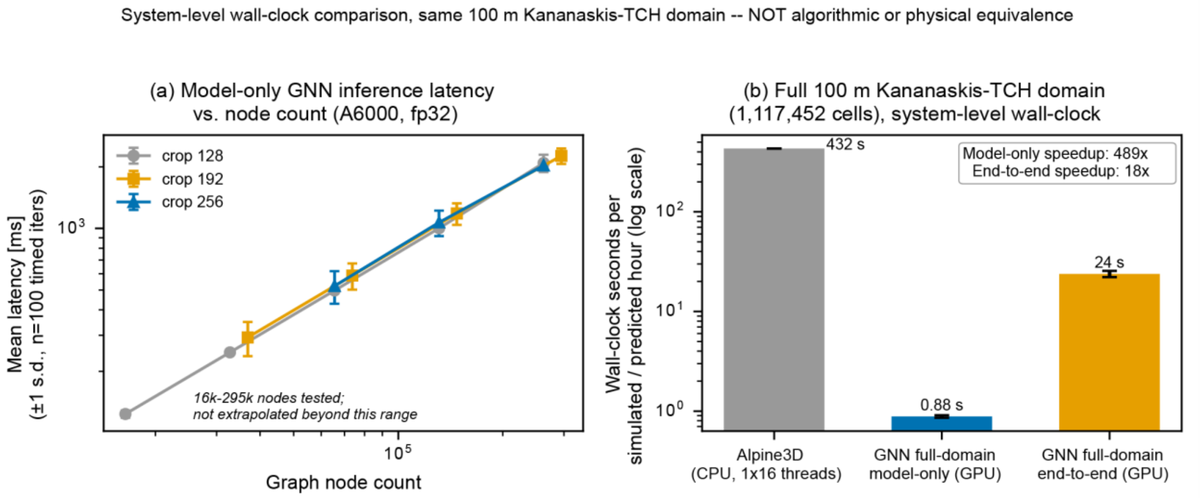

Figure 3: Runtime benchmark. Left: mean per-hour wall-clock time for GNN model-only inference, GNN end-to-end inference, and Alpine3D production simulation on the same 1.12-million-cell domain. Right: measured GNN latency across tested graph sizes. Hardware and functional scope differ, so values are system-level rather than hardware-normalized.

4. DISCUSSION

4.1 Persistence and surrogate evaluation

The persistence test changes how the R^2 values should be interpreted. Values near 0.999 mainly show that the current state is close to the state from the previous hour. This problem can occur whenever a slowly changing state is predicted over a short time step and the previous state is available as an input. For this reason, the surrogate needs to be evaluated against persistence and against the hourly changes that matter for the intended application.

The flux reconstruction tests whether the network learned selected process tendencies represented by the physics teacher. The positive holdout skill shows useful one-step learning. There is still a limitation with this test. The teacher-flux reconstruction does not close the state budget exactly at the hourly output cadence. It is a diagnostic and not a complete conservation guarantee. The hourly change distribution is also concentrated near zero. This creates a continuous-target imbalance that can favor the dominant regimes unless the rare but important changes are tested separately (Yang et al., 2021).

The corrected-2022 evaluation strengthens the persistence-based interpretation. Both flux-reconstructed quantities retained positive skill across a substantially larger corrected panel, while the absolute HS prediction remained persistence dominated. This shows that the process-variable evaluation is not limited to the original 2023 panel. The result should still be interpreted as robustness to a corrected forcing lineage within the same domain, not as independent temporal or geographic generalization.

4.2 Current operational limits

The model performs best during melt and runoff. The performance remains positive but is lower during accumulation and at low elevations. These regimes are relevant to snowpack interpretation, but the evaluation does not include avalanche occurrence, stability tests, snow profiles, or independent field observations. The test is not an avalanche-forecasting validation. The model was also not iterated over multiple days, so long-term stability has not been demonstrated.

The foundation model does not use wind as an input, and the teacher does not include wind transport. The model cannot reproduce wind-driven loading or erosion. Zero-shot tests at out-of-domain sites show that the architecture can execute on new terrain. However, useful flux-reconstructed skill did not clearly transfer. Geographic transfer and operational deployment have not been demonstrated.

The measured 18.1× end-to-end speedup is useful for repeated scenario analysis with the present foundation configuration. The larger 489× model-only difference shows that data loading and output transfer take more time than the network execution in the current workflow. The comparison uses different hardware and different functional scope. For this reason, it is a workflow benchmark for the SIMPLE-radiation, no-transport teacher. It is not a normalized comparison of equal physical functionality, and it does not measure performance for complex-radiation and transport-coupled Alpine3D configurations.

4.3 Radiation and transport testing plan

The future physics workflow is intended to evaluate complex terrain radiation and wind-driven redistribution separately. Controlled simulations allow each process to be studied without changing the foundation result reported in this paper.

The current extension is a proof of concept. One paired season was completed at one site, but parts of the simulation corpus require regeneration or re-verification. The transport-coupled simulation has not been compared with field observations, and continued training has not produced a converged and reproducible improvement. For these reasons, the transport and radiation work is included as preliminary future work rather than a main result.

Future work will require verified simulation data, comparison with appropriate baselines, and the same persistence-referenced evaluation used for the foundation model. Any radiation or transport result will be reported separately after it has been reproduced and validated.

4.4 Limitations and next steps

There are several limitations to the results. All reported skill is relative to a physics teacher and not to field truth. The foundation teacher uses simple radiation and does not include transport. Station-derived forcing for a subset of network stations also had a time zone labeling defect that has since been corrected. The corrected 2022 evaluation shows that the frozen checkpoint can retain positive flux-reconstructed skill after the forcing correction, but it does not remove the training-lineage limitation because the 2022 season was represented in the original training corpus.

The defect affected the mapping between observed weather and the simulated time of day.

The model is evaluated against the internally consistent teacher state and not against external real-world timestamps. This means the teacher-reproduction metrics remain valid, but the absolute real-world diurnal timing is provisional for the affected station fraction. The frozen checkpoint has only been tested one step at a time, and geographic transfer has not been established. The runtime comparison uses different hardware and does not have complete functional equivalence. Lastly, the paired radiation and transport corpus must be regenerated or re-verified before quantitative head results can be reported.

The next stage of the research will evaluate corrected multi-season simulations, higher-resolution domains, complex terrain radiation, and wind-driven redistribution. Long autoregressive rollouts, comparison with field observations, and function-matched compute-cost comparisons with Alpine3D are still required before operational use can be considered.

5. CONCLUSION

In this paper, a graph neural network surrogate was trained and tested on Alpine3D/SNOWPACK teacher simulations. The model learned useful one-step snowpack tendencies, but the performance was only clear after it was compared with persistence. On the fully held-out season, the copy-previous-hour baseline produced a higher R^2 than the absolute HS and SWE predictions, even though the model values were near 0.999. The changes calculated from the absolute-state predictions were also worse than persistence. Reconstructing the changes from the predicted flux and height-change variables produced positive skill of 0.373 for ΔHS and 0.795 for ΔSWE. A separate frozen-checkpoint evaluation on corrected 2022 simulations produced positive skill of 0.479 for ΔHS and 0.461 for ΔSWE across 290 stratified timesteps and approximately 312 million valid cell-time observations. This supports robustness to the corrected forcing lineage within the same domain, but it does not establish a new temporal or geographic holdout.

The strongest performance occurred during melt and runoff. On the full 1.12-million-cell domain, the end-to-end pipeline was 18.1× faster than the measured Alpine3D production workflow, while the model execution alone was 489× faster.

The results demonstrate a modelling and evaluation method and do not demonstrate an operational forecasting system. The foundation teacher does not include complex terrain radiation or wind transport. The surrogate has also not been tested through long rollouts, across different regions, or against field observations. Higher-fidelity radiation and transport extensions remain preliminary and should be reported only after they have been generated, reproduced, and validated.

REFERENCES

Bartelt, P. and Lehning, M., 2002. A physical SNOWPACK model for the Swiss avalanche warning, Part I: numerical model. Cold Regions Science and Technology, 35, 123–145. https://doi.org/10.1016/S0165-232X(02)00074-5.

Bavay, M. and Egger, T., 2014. MeteoIO 2.4.2: a preprocessing library for meteorological data. Geoscientific Model Development, 7, 3135–3151. https://doi.org/10.5194/gmd-7-3135-2014.

Charbonneau, A., Deck, K., and Schneider, T., 2025. A physics-constrained neural differential equation framework for data-driven snowpack simulation. Artificial Intelligence for the Earth Systems, 4(3). https://doi.org/10.1175/AIES-D-24-0040.1.

Helbig, N. and Löwe, H., 2012. Shortwave radiation parameterization scheme for subgrid topography. Journal of Geophysical Research: Atmospheres, 117, D03112. https://doi.org/10.1029/2011JD016465.

Lehning, M., Völksch, I., Gustafsson, D., Nguyen, T. A., Stähli, M., and Zappa, M., 2006. ALPINE3D: a detailed model of mountain surface processes and its application to snow hydrology. Hydrological Processes, 20, 2111–2128. https://doi.org/10.1002/hyp.6204.

Mott, R., Schirmer, M., Bavay, M., Grünewald, T., and Lehning, M., 2010. Understanding snow-transport processes shaping the mountain snow-cover. The Cryosphere, 4, 545–559. https://doi.org/10.5194/tc-4-545-2010.

Murphy, A. H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review, 116, 2417–2424. https://doi.org/10.1175/1520-0493(1988)116%3C2417:SSBOTM%3E2.0.CO;2.

Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A., and Battaglia, P. W., 2021. Learning mesh-based simulation with graph networks. International Conference on Learning Representations. https://openreview.net/forum?id=roNqYL0_XP.

Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., and Battaglia, P. W., 2020. Learning to simulate complex physics with graph networks. Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 8459–8468. https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Yang, Y., Zha, K., Chen, Y.-C., Wang, H., and Katabi, D., 2021. Delving into deep imbalanced regression. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 11842–11851. https://proceedings.mlr.press/v139/yang21m.html.