Uncertainty-Aware Learning With Label Noise for Glacier Mass Balance Modeling

Codrut-Andrei Diaconu, *Graduate Student Member, IEEE*, and Nina Maria Gottschling

***Abstract*— Glacier mass balance (MB) modeling is crucial for understanding the impact of climate change on Earth's freshwater resources and sea-level rise. Recent works have shown the benefit of using machine learning (ML) and deep learning (DL) methods to better capture the nonlinearities in the system than commonly used temperature-index models. However, when relying on remote sensing products for training, the presence of data noise is a challenge for these methods, and therefore quantifying the uncertainty becomes essential. In this work, we produce a tabular dataset consisting of annual MBs for 1000 glaciers over 20 years with meteorological and topographical input features. Using this dataset, we systematically study various uncertainty estimation methods and their impact on the quality of the predictions. Our experimental results show that ensemble methods are promising for capturing the uncertainty in the data: their predictions are more accurate, more robust against label noise, and better calibrated. In particular, the multilayer perceptron (MLP) ensemble coupled with an explicit noise model shows an increase of up to 5.5% in the explained variance and is much less affected by the gradually injected label noise: the average mean absolute error (MAE) increases at a rate twice smaller. For reproducibility, code and data are available at https://github.com/dcodrut/oggm_smb_dl_uq.**

***Index Terms*— Ensemble learning, glacier mass balance (MB) modeling, noisy labels, robustness, uncertainty quantification (UQ).**

I. Introduction

THE cryosphere, as any other component of the Earth system, is highly complex and nonlinear. Modeling it accurately remains challenging, especially at regional scale [1]. As the societal and environmental impact of the retreat of glaciers is certain [2], appropriate methods for modeling and predicting the evolution of the glaciers are important to adapt necessary policies [3]. The glacier mass balance (MB), defined as the sum of accumulation (e.g., through snow, avalanches, refreezing of rain) and ablation (e.g., through surface melting, drifting snow, sublimation) [4], over a fixed period of time, is one of the most important components in glacier modeling and also one of the essential climate variables (ECVs) [5].

In [6], it was shown that there is a significant nonlinear part in the relationship between climate and MB. Supporting this assumption, Bolibar et al. [7] show that deep learning (DL) captures the nonlinear response of MBs to temperature and precipitation, especially in extreme cases, better than classical approaches, such as linear statistical and temperature-index models.

This is opposed to the commonly used glacier MB models that can be applied at a large scale. These often rely on temperature-index models [1], which assume a linear relationship between the days with above zero temperatures and the melting of ice or snow [8]. Hence, it is promising to apply DL methods, such as nonlinear neural networks (NNs) or classical machine learning (ML) models, such as random forest (RF), as a statistical method to predict glacier MBs. However, these models use data based on in situ measurements (e.g., using ablation stakes) or remote-sensed data (e.g., using digital elevation model (DEM) differencing) [9], [10]. Both the approaches have nonnegligible uncertainties due to measurement errors, sampling biases, or shortcomings in the methodology. For example, the in situ measurements have an accuracy typically lying between 0.1-m water equivalent (w.e.) and 0.6-m w.e. [11]. Thus, MB models are trained with noisy labels, yet should ideally make noise-free predictions. Uncertainty quantification (UQ) methods could solve this issue, by modeling the data noise and the model uncertainty, and thereby disentangling them from the mean predictions.

ML has recently become popular for MB modeling: [7] projects the 21st-century glacier evolution in the French Alps with a standard multilayer perceptron (MLP) model for MB as a better alternative to linear regression (LR); [12] models winter point MBs using gradient boosting regressor (GBR); [13] estimates annual point MBs using four different methods, i.e., support-vector machine (SVM), RF, GBR, and MLP. None of these studies models any uncertainty source and only uses the testing errors as a quality indicator. Given that glaciers are critical components in the Earth system and a significant percentage of the world's population (∼22%) is relying on their water storage capacity [2], if policy makers are to make decisions based on predictions obtained from ML/DL methods, then it is paramount that they are not just a black-box tool but provide reliable uncertainty estimates. We aim to bridge this gap, by coupling NNs with different UQ methods for MB prediction and investigating their behavior with respect to label noise, by making the following] contributions.

Manuscript received 29 September 2023; revised 22 December 2023; accepted 3 January 2024. Date of publication 19 January 2024; date of current version 6 February 2024. This work was supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@FZJ Partition. The work of Codrut-Andrei Diaconu was supported by the Helmholtz Association through the Joint Research School Munich School for Data Science–(MuDS) under Grant HIDSS-0006. *(Corresponding author: Codrut-Andrei Diaconu.)*

Codrut-Andrei Diaconu is with the German Aerospace Center (DLR), 82234 Weßling, Germany, and also with the Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: codrut-andrei.diaconu@dlr.de).

Nina Maria Gottschling is with the German Aerospace Center (DLR), 82234 Weßling, Germany.

This article has supplementary downloadable material available at https://doi.org/10.1109/LGRS.2024.3356160, provided by the authors.

Digital Object Identifier 10.1109/LGRS.2024.3356160

1) We provide a dataset for MB regression suitable for studying UQ methods.
2) By systematically adding label noise, we compare various models (LR, RF, and six MLP versions coupled with different UQ components) with respect to their predictive performance, the quality of the UQ estimates, and their robustness against noisy labels.

II. Dataset

There are various limitations for datasets of MB reconstructions. For in situ measurements, these include limited annual glacier-wide observations, e.g., the world glacier monitoring service (WGMS) [14]—a database gathering all in situ measurements—contains less than 500 glaciers, which is considerably less than the almost 200 000 glaciers worldwide [15]. In addition, there are uncertainties such as measurement accuracy, the distribution of the ablation stakes or snow pits, and the interpolation method, which are glacier-specific and difficult to estimate [11]. For MB estimates based on remote sensing techniques, an advantage is the increased coverage, and various approaches to estimate glacier MB have been proposed (see Table 3 from [10]). However, there are also sources of uncertainties (e.g., the volume-to-mass conversion) and discrepancies between derived MB estimates [10]. In addition, geodetic estimates at glacier level are usually available as multiannual averages, thus limiting their use for calibrating annual/seasonal models [10]. Thus, we use MB reconstructions instead of measurements to investigate the potential of ML and DL models for annual glacier MB modeling.

A. Dataset Construction

We use the open global glacier model (OGGM) [16], an open-source framework for glacier modeling, to reconstruct the annual MBs of the 1000 largest glaciers in Central Europe (out of 3927, cf. [15]), which cover about 90% of the total glaciated area in the region. The MB model used in OGGM requires temperature and precipitation as drivers [17] which are obtained from [18]. OGGM calibrates its parameters using the 20 years average MBs from [19]. We limit the analysis to the same 20-year period (i.e., $1999-2019$), resulting in a total of $20 * 1000 = 20\,000$ data entries. As inputs, we use the same meteorological drivers (i.e., monthly temperature and precipitation averages) as OGGM, as well as six topographical features (area, minimum, maximum and mean elevation, slope, aspect—sine & cosine), resulting in a 31-D input. The topographical features are added to compensate for the fact that OGGM estimates the MB in a pointwise manner, along multiple lines distributed over a glacier (called flow lines [16]), whereas in our approach we train a glacierwide MB model.

B. Label Noise Injection

The reconstructed MBs have a mean of -0.73-m w.e, reflecting the observed mass loss over the past two decades [19], and a standard deviation $\sigma_{\text{data}} \approx 0.78$-m w.e. We inject Gaussian noise in the labels and build five scenarios denoted by $z = z_{\text{noise}}$, $z_{\text{noise}} \in \{0.1, 0.2, \ldots, 0.5\}$, where z_{noise} controls the noise values η relative to σ_{data}: $\eta \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$ where $\sigma_{\text{noise}} = z_{\text{noise}} \cdot \sigma_{\text{data}}$. This results in a noise standard deviation varying from 0.08- to 0.39-m w.e., similar to the range of the errors estimated for the in situ measured data [11]. Noise-free labels are denoted by $z = 0.0$. Given that the region we cover in our dataset is relatively small, we found that a Gaussian homoscedastic noise model is a reasonable choice, as it approximately matches the estimated errors of the observed MBs used for calibrating OGGM [19]. A more detailed explanation and limitations of this choice are provided in the Supplementary. Another reason is that the focus of our study is on investigating whether coupling the models with UQ components improves their robustness against label noise, making use of the total predictive uncertainties rather than focusing on the aleatoric uncertainty alone.

III. Methods

A. Brief Introduction of Methods

Given the set of input–target pairs from our dataset, $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{K}\}$, the task of the models is to predict a target $y^\star \in Y$ given an input $x^\star \in X$ such that the loss objective between the predictions and targets is minimized over all the training points. The model can be regarded as a function f_θ, parameterized by weights θ, which maps inputs x directly to targets $y \in Y$, $f_\theta : X \to Y$ or to a probability distribution, $f_\theta : X \to \mathcal{P}(Y)$ such that $f_\theta(x^\star) = p_\theta(x^\star) \in \mathcal{P}(Y)$.

In the following, we briefly describe the eight models used, which include an LR model, an RF Regressor, and five variants of NNs built upon an MLP. A more detailed description is provided in the supplementary material.

Linear Regression: standard multi-LR model used as baseline, to support the claim that nonlinear models are more suitable for glacierwide MB modeling.

Random Forest: introduced by Breiman [20], it consists of training randomized decision trees using bootstrapping and then aggregate the predictions by averaging them. A review of RF as a powerful tool for classification and regression is provided in [21]. We moreover consider the variance of the predictions as a measure of predictive uncertainty.

Multilayer Perceptron: a simple fully connected network with two hidden layers and a nonlinear activation function, used as baseline.

Gaussian MLP (MLP+NLL): a deterministic model that predicts the parameters of a Gaussian distribution

$$f_\theta(x^\star) = (\mu_\theta(x^\star), \sigma_\theta(x^\star)) \tag{1}$$

in a single forward pass, where standard deviations $\sigma_\theta(x^\star)$ can be used as a measure of data uncertainty. This is achieved by extending the output of the previous architecture to two dimensions and train it with the negative log-likelihood (NLL) of a Gaussian as a loss objective [22].

MC-Dropout (MLP+MCD): an approximate Bayesian method with sampling, as in [23]. A fixed dropout rate p is added, meaning that the weights are randomly set to zero during each forward pass with the probability p. This models the network weights and biases as a Bernoulli distribution with dropout probability p. We also consider combining this method with the previous model (Gaussian MLP), as in [22], aiming for disentangling the data and model uncertainties, abbreviated as **MLP+NLL+MCD**.

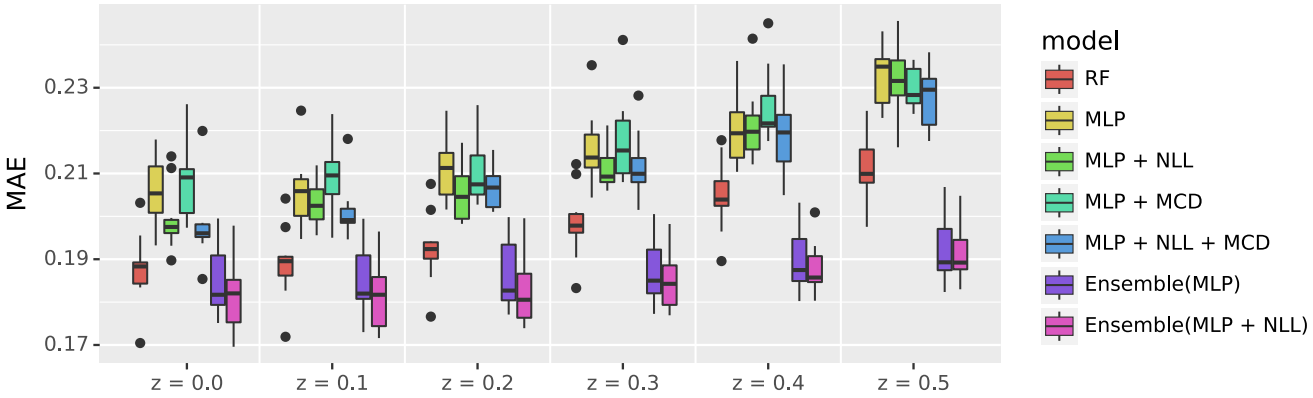

Fig. 1. **Robustness evaluation:** Test performance (MAE) for all the models (except LR) trained on multiple levels of noise and evaluated on clean labels.

TABLE I

ACCURACY METRICS: PERFORMANCE SCORES ($\mu \pm \sigma$) EVALUATED ON CLEAN LABELS FOR THE MODELS TRAINED WITH AN AVERAGE AMOUNT OF NOISE ($z = 0.3$)

model	MAE ↓	RMSE ↓	R^2 ↑
LR	0.385 ± 0.007	0.497 ± 0.017	0.590 ± 0.017
RF	0.198 ± 0.008	0.272 ± 0.026	0.876 ± 0.021
MLP	0.216 ± 0.009	0.297 ± 0.026	0.853 ± 0.025
MLP+NLL	0.211 ± 0.005	0.285 ± 0.020	0.864 ± 0.017
MLP+MCD	0.218 ± 0.010	0.293 ± 0.028	0.856 ± 0.026
MLP+NLL+MCD	0.212 ± 0.008	0.285 ± 0.022	0.865 ± 0.019
Ensemble(MLP)	0.187 ± 0.007	0.257 ± 0.027	0.889 ± 0.022
Ensemble(MLP+NLL)	**0.185 ± 0.007**	**0.252 ± 0.024**	**0.894 ± 0.019**

Deep Ensembles [Ensemble (MLP)]: introduced in [24], Deep Ensembles approximate a posterior distribution over the model weights with a Gaussian mixture model over the output of separately initialized and trained networks. Wilson and Izmailov [25] showed that Deep Ensembles can be interpreted as a Bayesian method. In addition, each ensemble member can be a Gaussian MLP, denoted as **Ensemble (MLP+NLL)**.

B. Metrics

Regression tasks are commonly evaluated by accuracy metrics such as root mean squared error (RMSE), mean absolute error (MAE), or coefficient of determination (R^2). A better quality of prediction is indicated by a lower RMSE and MAE and an R^2 score close to 1.0. However, these measures only characterize the error between point predictions and available targets. To compare the predictive uncertainties to the target distribution, we need additional metrics, such as proper scoring rules [26]. We consider the NLL of a Gaussian as a proper scoring rule [26]. We also report the miscalibration area, where a lower miscalibration area indicates a better fit of the predictive uncertainties to the true target distribution. To quantify the overall confidence of a model in a single metric, we consider sharpness which computes the mean of the predictive uncertainties. We use [27] for computing these metrics.

IV. EXPERIMENTAL RESULTS

A. Evaluation Details

From the dataset, we keep 20% of the glaciers for testing and the remaining are split at glacier level into training and validation (90% and 10%, respectively). To reduce the impact of randomness in our results, we repeat the experiments ten times with different data splits and different model initializations. The hyperparameters of each method are provided in the Supplementary.

B. Evaluation of Mean Predictions

Controlling the label noise allows us to compare the models with respect to robustness, by analyzing which models can still predict the true labels accurately when trained with increasing label noise.

In Table I, we show the MAE, RMSE, and R^2 on clean labels for all the models trained with the noisy labels with $z = 0.3$. LR performs the worst with a significant gap compared with the other models. The two MLP ensemble versions perform the best on all the metrics, closely followed by RF. For R^2 scores, we observe that all the methods (except LR) attain values in [85%, 90%], the Ensemble (MLP+NLL) model outperforming RF only by 1.8%. The tables with the accuracy metrics for the other noise levels are included in the supplementary material and show the same trends. We also included the mean bias error (MBE) as an additional metric, which is in general very small (less than 3-cm w.e.), with little variance across methods and no correlation to the noise level.

We investigate the robustness to training on increasing label noise by assessing the MAE of the models. Fig. 1 shows the MAE distribution of the ten differently initialized models. Taking into account the variance due to initialization, the two MLP ensemble methods still perform best, followed by RF. As expected, the MAE increases with increasing label noise for all the models. Table I shows that the variance of the results is comparable across methods.

To assess which models are affected the most by the increasing noise, we show the average MAE scores obtained when training on clean labels as a baseline and compute the change (expressed in percentages) when training on noisy labels in Table II. All the models show increasing MAE with increasing noise, reaching up to 16.5% increase (MLP+NLL for $z = 0.5$). The Ensemble (MLP) and Ensemble (MLP+NLL) increase only by 3.6% and 5.4%, respectively, whereas the others exceed +10%. The models trained with NLL (i.e., MLP+NLL and MLP+NLL+MCD) are relatively more affected, reaching +10% already at $z = 0.4$. This is also

TABLE II

RELATIVE PERFORMANCE DIFFERENCE: CHANGE IN AVERAGE MAE SCORES, EVALUATED ON CLEAN LABELS, WHEN TRAINING ON NOISY LABELS ($z \geq 0.1$) COMPARED WITH TRAINING ON CLEAN LABELS ($z = 0.0$)

model	z = 0.0	z = 0.1	z = 0.2	z = 0.3	z = 0.4	z = 0.5
RF	0.188	+0.6%	+2.5%	+5.7%	+9.0%	+12.5%
MLP	0.206	**-0.2%**	+2.4%	+4.8%	+7.1%	+13.1%
MLP+NLL	0.199	+1.8%	+2.8%	+6.0%	+11.0%	+16.3%
MLP+MCD	0.209	+0.1%	+0.6%	+4.3%	+8.1%	+10.1%
MLP+NLL+MCD	0.198	+1.8%	+4.7%	+7.3%	+10.8%	+15.3%
Ensemble(MLP)	0.186	-0.0%	**+0.5%**	**+1.1%**	**+2.3%**	**+3.6%**
Ensemble(MLP+NLL)	0.181	-0.0%	+0.6%	+1.8%	+3.6%	+5.4%

reflected when comparing the two MLP Ensemble versions, where Ensemble (MLP) performs better.

C. Evaluation of Predictive Uncertainties

In Section IV-B, we showed that coupling the models with uncertainty estimation components helps improve their robustness and yields improved accuracy. Yet, in many applications it also important to provide uncertainty estimates, e.g., to make risk assessments or withdraw from predictions that have high uncertainty. In a real-world scenario, one does not have access to clean labels which we previously exploited for robustness evaluation. In this section, we investigate how well the UQ methods capture the uncertainties, using the metrics described in Section III-B evaluated with the noisy labels. Moreover, we assess whether these methods provide useful uncertainties using selective prediction, as introduced in [28]. Here, samples with a predictive uncertainty above a given threshold are omitted from prediction and, e.g., referred to an expert or a different method. If larger uncertainties are correlated with worse predictions, this increases overall accuracy.

Table III shows miscalibration area, sharpness, and NLL scores, obtained for the average noise case ($z = 0.3$). Compared with the previous results, discrepancies between the methods are higher. The Ensemble (MLP+NLL) model obtains a lower miscalibration area compared with Ensemble (MLP) (which performs the worst), closely followed by RF. The sharpness is much smaller for the standard MLP ensemble. There are large variations for the NLL scores and the Ensemble (MLP+NLL) performs again the best, with RF performing similarly and Ensemble (MLP) the worst. The tables with the UQ metrics for the other noise levels are included in the supplementary material and show the same trends. For a more detailed analysis, a figure of the calibration curves for all the noise levels is also included in the supplementary, where it can be observed that the Ensemble (MLP) model is highly overconfident, reflected by the high miscalibration area in Table III.

Finally, we assess whether the uncertainty scores can improve the accuracy with selective prediction. Fig. 2 shows the average performance (MAE), of each model against the coverage percentage, i.e., the percentage of samples with the lowest predictive uncertainties, the remaining ones being dropped. Ideally, we want to see a better performance when using the least uncertain samples but we can see that only the ensemble methods (including RF) have this behavior.

TABLE III

UQ METRICS: ($\mu \pm \sigma$) FOR THE MODELS TRAINED AND EVALUATED ON THE NOISY LABELS WITH $z = 0.3$

model	miscalibration area ↓	sharpness ↓	NLL ↓
RF	0.014 ± 0.009	0.363 ± 0.020	0.406 ± 0.045
MLP+NLL	0.055 ± 0.015	0.314 ± 0.020	0.545 ± 0.106
MLP+MCD	0.164 ± 0.027	0.210 ± 0.017	1.086 ± 0.261
MLP+NLL+MCD	0.029 ± 0.010	0.402 ± 0.010	0.425 ± 0.035
Ensemble(MLP)	0.260 ± 0.013	**0.154 ± 0.018**	3.698 ± 0.603
Ensemble(MLP+NLL)	**0.008 ± 0.003**	0.348 ± 0.012	**0.349 ± 0.045**

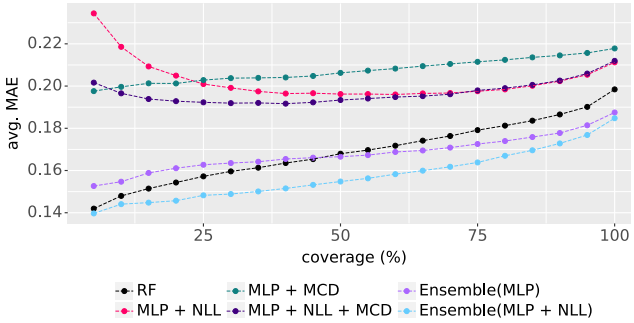

Fig. 2. **Selective prediction:** Test performance (MAE) on clean labels averaged for all the data points which have the estimated total uncertainty score below a certain threshold (x-axis). The models are trained on the noisy labels with $z = 0.3$.

Selective prediction applied to the models trained on the other noise levels shows similar trends and is included in the supplementary material.

V. DISCUSSION

The results described in Section IV-B indicate that a linear model is not sufficient for glacier-wide MB modeling, thus suggesting that the problem is nonlinear, as found in previous studies [6], [29]. Among the nonlinear methods, we observe that the top performing ones are the ensemble methods, including RF. The fact that Ensemble (MLP+NLL) yields the overall best results provides evidence that coupling the model with this aleatoric uncertainty component also improves predictions. However, training a RF remains easier and faster, which makes it also a good candidate, with a relatively small performance gap compared with Ensemble (MLP+NLL). We also found RF to be less sensitive to the choice of hyperparameters compared with the MLPs; probably also explained by the larger ensemble size (up to 500 trees were used versus only ten for the MLPs).

When analyzing the influence of increasing the label noise on performance, ensembles of MLPs are again favored, as their performance degrades slower compared with the other methods. Here, the gap between RF and the ensembles of MLPs is higher, which indicates RF is more prone to overfitting on our dataset. In the large-scale study from [30], tree-based models were found to perform better on tabular data than NNs, as they can approximate irregular functions whereas NNs tend to be biased toward smoother solutions. However, in our context, this inductive bias could be beneficial when dealing with large amounts of noise, potentially making NNs less prone to overfitting, an aspect which was previously investigated for classification tasks in [31].

Concerning UQ (Section IV-C), the complementary metrics we used (i.e., calibration, sharpness and NLL) reveal that Ensemble (MLP+NLL) matches our dataset distribution

the best. Ensemble (MLP) is overconfident, which explains the comparably low sharpness. This indicates that predicting the parameters of a Gaussian enables disentangling the model and data uncertainty, shown in the figures of model and data uncertainty in the supplementary material. The RF also provides relatively well calibrated predictions. Furthermore, the selective prediction results also indicate that the three ensemble methods perform the best. The Ensemble (MLP+NLL) is slightly outperforming, as its average MAE stays relatively low until it reaches a significant coverage (>= 75%), thus making it a good candidate in practice. The nonensemble models perform in general worse, both from the perspective of predictive power and uncertainty estimation.

VI. Conclusion

We introduce a simple and relatively small dataset for an important regression task in glacier modeling: predicting the annual MB using meteorological and topographical drivers. We then compare various methods (LR, RF, MLPs, and ensembles of MLPs) on how they perform when trained with noisy labels while still evaluating them on the clean labels. The ensemble methods performed the best (including RF), being more robust when increasing the label noise. When coupling the ensemble of MLPs with a Gaussian output, thus explicitly modeling the data uncertainty, the performance increases and the predictions become significantly better calibrated. The uncertainties from the ensemble methods can also be used for selective prediction, leading to more accurate and reliable MB predictions while still keeping a significant coverage.

We would therefore recommend the ensemble methods for glacier-wide MB modeling to the cryosphere community, in particular the Ensemble (MLP+NLL) version. However, these models are sensitive to the hyperparameters, so significant effort should be allocated to tuning these. From this perspective, RF was more robust but we would still recommend performing HPO: we observed that the final models grow smaller trees when having a large amount of noise, an indicator that HPO can prevent overfitting.

One promising extension of this study is to inject noise in the input data based on certain features (i.e., a heteroscedastic noise model) which is closer to the real setup, as for instance, the remote-sensing-based MBs from [19] have larger errors for small glaciers.

References

[1] B. Marzeion et al., "Partitioning the uncertainty of ensemble projections of global glacier mass change," *Earth's Future*, vol. 8, no. 7, Jul. 2020, Art. no. e2019EF001470.

[2] W. W. Immerzeel et al., "Importance and vulnerability of the world's water towers," *Nature*, vol. 577, pp. 364–369, Dec. 2020.

[3] D. R. Rounce et al., "Global glacier change in the 21st century: Every increase in temperature matters," *Science*, vol. 379, no. 6627, pp. 78–83, Jan. 2023.

[4] J. G. Cogley et al., "Glossary of glacier mass balance and related terms," Int. Hydrol. Programme, Tech. Rep., 2010. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000192525

[5] S. Bojinski, M. Verstraete, T. C. Peterson, C. Richter, A. Simmons, and M. Zemp, "The concept of essential climate variables in support of climate research, applications, and policy," *Bull. Amer. Meteorol. Soc.*, vol. 95, no. 9, pp. 1431–1443, Sep. 2014.

[6] D. Steiner, A. Walter, and H. J. Zumbühl, "The application of a non-linear back-propagation neural network to study the mass balance of grosse aletschgletscher, Switzerland," *J. Glaciol.*, vol. 51, no. 173, pp. 313–323, 2005.

[7] J. Bolibar, A. Rabatel, I. Gouttevin, H. Zekollari, and C. Galiez, "Nonlinear sensitivity of glacier mass balance to future climate change unveiled by deep learning," *Nature Commun.*, vol. 13, no. 1, p. 409, Jan. 2022.

[8] R. Hock, "Temperature index melt modelling in mountain areas," *J. Hydrol.*, vol. 282, nos. 1–4, pp. 104–115, Nov. 2003.

[9] J. Graham Cogley, "Geodetic and direct mass-balance measurements: Comparison and joint analysis," *Ann. Glaciol.*, vol. 50, no. 50, pp. 96–100, 2009.

[10] E. Berthier et al., "Measuring glacier mass changes from space—A review," *Rep. Prog. Phys.*, vol. 86, no. 3, 2023, Art. no. 036801.

[11] M. Zemp, M. Hoelzle, and W. Haeberli, "Six decades of glacier mass-balance observations: A review of the worldwide monitoring network," *Ann. Glaciol.*, vol. 50, no. 50, pp. 101–111, 2009.

[12] M. Guidicelli, M. Huss, M. Gabella, and N. Salzmann, "Spatio-temporal reconstruction of winter glacier mass balance in the alps, Scandinavia, central Asia and Western Canada (1981–2019) using climate reanalyses and machine learning," *Cryosphere*, vol. 17, no. 2, pp. 977–1002, Mar. 2023.

[13] R. Anilkumar, R. Bharti, D. Chutia, and S. P. Aggarwal, "Modelling point mass balance for the glaciers of the central European Alps using machine learning techniques," *Cryosphere*, vol. 17, no. 7, pp. 2811–2828, Jul. 2023.

[14] WGMS, *Fluctuations of Glaciers Database*, World Glacier Monitoring Service (WGMS), Zurich, Switzerland, 2022, doi: 10.5904/wgms-fog-2022-09.

[15] W. T. Pfeffer et al., "The Randolph glacier inventory: A globally complete inventory of glaciers," *J. Glaciol.*, vol. 60, no. 221, pp. 537–552, 2014.

[16] F. Maussion et al., "The open global glacier model (OGGM) v1. 1," *Geosci. Model Develop.*, vol. 12, no. 3, pp. 909–931, 2019.

[17] B. Marzeion, A. H. Jarosch, and M. Hofer, "Past and future sea-level change from the surface mass balance of glaciers," *Cryosphere*, vol. 6, no. 6, pp. 1295–1322, Nov. 2012.

[18] M. Cucchi et al., "WFDE5: Bias-adjusted ERA5 reanalysis data for impact studies," *Earth Syst. Sci. Data*, vol. 12, no. 3, pp. 2097–2120, Sep. 2020.

[19] R. Hugonnet et al., "Accelerated global glacier mass loss in the early twenty-first century," *Nature*, vol. 592, no. 7856, pp. 726–731, Apr. 2021.

[20] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[21] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, Jun. 2016.

[22] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. NeurIPS*, 2017, pp. 5574–5584.

[23] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.

[24] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. NeurIPS*, 2017, pp. 6402–6413.

[25] A. G. Wilson and P. Izmailov, "Bayesian deep learning and a probabilistic perspective of generalization," in *Proc. NeurIPS*, 2020, pp. 4697–4708.

[26] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, Mar. 2007.

[27] Y. Chung, I. Char, H. Guo, J. Schneider, and W. Neiswanger, "Uncertainty toolbox: An open-source library for assessing, visualizing, and improving uncertainty quantification," 2021, *arXiv:2109.10254*. [Online]. Available: https://arxiv.org/abs/2109.10254

[28] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Proc. NeurIPS*, 2017, pp. 4878–4887.

[29] J. Bolibar, A. Rabatel, I. Gouttevin, C. Galiez, T. Condom, and E. Sauquet, "Deep learning applied to glacier evolution modelling," *Cryosphere*, vol. 14, no. 2, pp. 565–584, Feb. 2020.

[30] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Proc. NeurIPS*, 2022.

[31] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," 2017, *arXiv:1705.10694*.