On Data Driven bias correction of operational Weather Forecasts in the High Arctic Marginal Ice Zone

Mats Ippach[1], **Are Frode Kvanum**[1,2]

[1]Department of Geosciences, University of Oslo
[2]Development Centre for Weather Forecasting, Norwegian Meteorological Institute

1 Introduction

In the forecasting of Arctic weather systems, the interaction between ocean, ice, and atmosphere plays a crucial role as it governs the exchange of energy at the respective interfaces. However, these interfaces are usually not discrete, but rather transitional zones, whose implementation in weather forecasting systems can be crucial. The Marginal Ice Zone (MIZ) forms such a transitional interface between the solid sea ice and the open ocean. To improve the understanding of the role of the MIZ in coupled Arctic forecasting, the Svalbard Marginal Ice Zone 2025 Campaign, set out to put an observational network of buoys in place, which monitors air and surface temperature spatio-temporally, among other parameters (Müller et al., 2025).

Due to the complexity of numerical weather prediction systems, systematic errors commonly arise due to geographical, physical or numerical limitations in the model. Batrak and Müller (2019) identified that numerical weather predictions systems commonly overestimate surface temperatures in the Arctic, especially during cold conditions, caused by a lack of snow on sea ice representation the models. However, improving model physics is usually a difficult task, and can easily cause unintended responses in other model components. Hence, statistical bias correction is commonly employed as a postprocessing step to improve prediction quality (Vannitsem et al., 2021). Recently, Hieta and Partio (2025) demonstrated that by applying an XGBoost machine learning model for bias correction, weather forecasts from the Finnish Meteorological Institute had their RMSE reduced by 17% - 43% depending on the variable and lead time. With similar methods, Palerme and Müller (2021) was able to reduce sea ice drift forecast error by 8%, on average improving 55.7% of considered forecasts.

In this adjacent course project, the non-quality-controlled data of selected monitoring buoys will be utilized along with air temperature data from the Arctic numerical weather prediction system AROME Arctic (Müller et al., 2017) and sea ice concentration from CARRA (Schyberg et al., 2020) to investigate the relationship between observed and forecast temperature. The novel SvalMIZ25 observational network compensates for the prior lack of empirical data Müller et al. (2025), on which we will employ various statistical models and explore the possibility of performing a statistical bias correction to improve the near-surface air temperature forecasts from AROME Arctic. To correct the hypothesized bias in the forecast, available parameters will be used to train a linear regression model and machine learning in form of a neural network and decision trees. A significant improvement of the forecast indicates the importance of bias correction in the MIZ for weather prediction systems.

2 Study area

The SvalMIZ-25 campaign deployed 21 OpenMetBuoys in the MIZ North-West of the Svalbard archipelago between 22.April - 11.May 2025. During the campaign, a large amount of sea ice was present around Svalbard. Most of the campaign was conducted during a cold phase with temperatures ranging below 0 to -10 degrees Celsius. Fig. 1 summarizes our data through time-series of the temperature and sea ice concentration

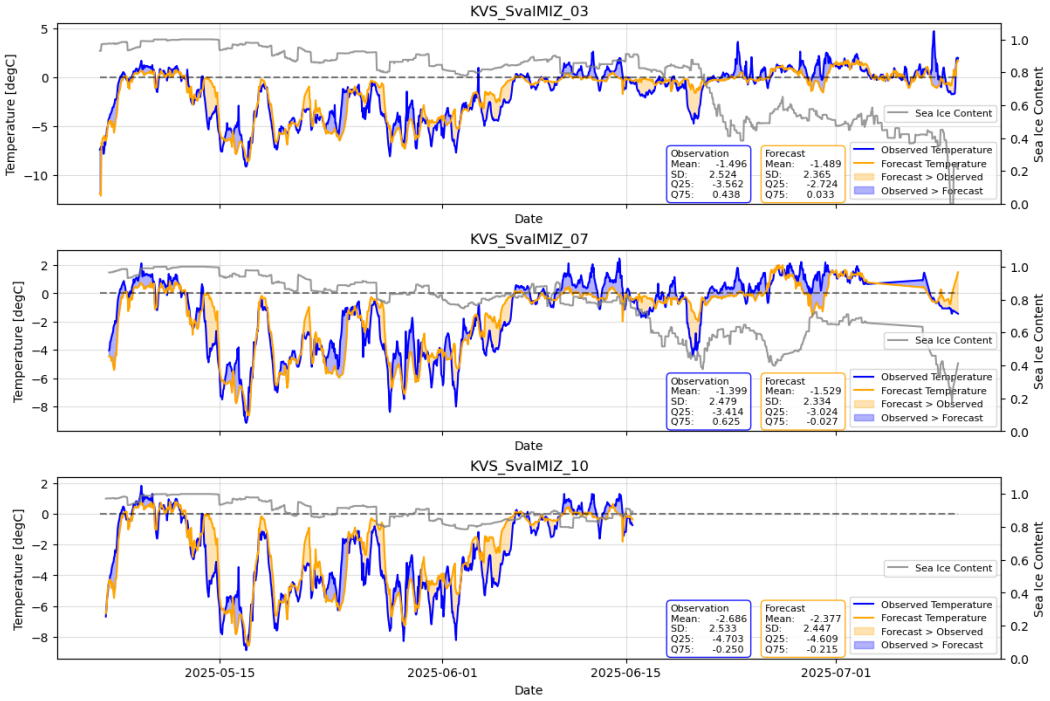

Figure 1. Time series of the observed air temperatures from buoys of the SvalMIZ25 observational network in the MIZ combined with spatially and temporally matched air temperature predictions from AROME Arctic and sea ice content (SIC) from CARRA. Shaded area indicate differences between the observed and forecast temperatures.

from three SvalMIZ-25 buoys. For further details, we refer to the cruise report by Müller et al. (2025).

3 Data

3.1 Dataset Description

We construct our dataset by selecting three buoys from the SvalMIZ25 dataset (03, 07 and 10) found in Müller et al. (2025). Due to the non-quality controlled state of the data, we threshold buoys based on lifespan and signs of faulty data. The selected buoys cover a period of 22.April - 10.July 2025, see Fig. 1.

We pair the buoy dataset with atmospheric forecasts fetched from the AROME Arctic numerical weather prediction system (Müller et al., 2017). AROME Arctic has a regional coverage, including our study area. Additionally, the system features a 2.5 km spatial and 1 hour temporal resolution, thus we expect the model to have the theoretical capabilities to capture the physical processes in the atmosphere of our region of interest. The gridded atmospheric forecast data is reprojected onto the buoy positions.

No direct observations of sea ice concentration (SIC) is present in the dataset. Thus, we leverage a SIC reanalysis which aims to supply a proxy of SIC. We use the SIC found in the Copernicus Regional Arctic Reanalysis (CARRA) (Schyberg et al., 2020), which is generated using AROME Arctic.

3.2 Data preparation

We separate the buoys into training and test folds, with buoy 03 and 07 constituting the training fold (2936 samples) and buoy 10 used for testing (963 samples). We further split 20% of the training fold into a validation fold when training appropriate models.

The full set of predictor variables used for training models includes estimated T2M from AROME Arctic, SIC from CARRA, as well as temporal predictors derived from the day of year (doy) and hour of day (hod). In order to conserve the periodic properties pertaining to both doy and hod, both predictors are decomposed into sine and cosine components.

Instead of directly predicting the observed temperature, we use as target the residuals between observed and estimated temperature. Thus, to obtain the corrected temperature value we add the predicted residual to the estimated temperature during inference.

4 Materials and Methods

4.1 Comparison of distributions

4.1.1 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (KS) is applied to test the similarity of two empirical distributions (two-sample KS test), i.e., whether the observation and forecast data originate from the same distribution. The KS test was also chosen because it doesn't require knowing the sample distribution.

The test is based on the comparison of the respective cumulative frequencies resulting from the cumulative distribution functions (CDF). The test statistic describes the absolute maximum deviation D_{max} of the frequencies and is compared to the critical value resulting from the significance level (α) and sample size. H_0 states that two samples originate from the same distribution and can be rejected if D_{max} is greater than the critical value.

Test statistic:

$$D_{max} = max|F_{1,n}(x) - F_{2,m}(x)|$$

Critical value:

$$c(\alpha) * \sqrt{\frac{n+m}{n*m}}$$

Hypotheses:

$$H_0: \quad F_{1,n}(x) = F_{2,m}(x) \qquad H_a: \quad F_{1,n}(x) \neq F_{2,m}(x)$$

Rejection criterion:

$$D_{max} > c(\alpha)\sqrt{\frac{n+m}{n*m}} \quad or \quad D_{max}\sqrt{\frac{n*m}{n+m}} > c(\alpha), \quad with \quad c(\alpha) = \sqrt{-\ln(\frac{\alpha}{2}) * \frac{1}{2}}$$

4.1.2 Model evaluation

To evaluate biases between the forecast and observed data, quantile-quantile plots (QQ-plots) and probability-probability plots (PP-plots) were used. QQ-plots show same quantiles of two sorted samples against another. These plots are used to identify biases in specific ranges (i.e., temperature ranges) with a higher sensitivity towards their tails. The PP-plots are constructed analogously, but consider the probability values derived from the empirical CDF. Complementary to the QQ-plots, the PP-plots show differences in the skewness of the sample distributions, whereas the sensitivity is centered around the mode.

4.1.3 Residuals and Skill Score evaluation

The residuals between the observed and estimated variable ($r_i = y_i - x_i$) will be used to evaluate a temporal bias and validate their distribution visually as Gaussian. Confirming the latter allows us to estimate the confidence intervals of the root mean square error (RMSE).

To quantify the forecast error and evaluate the bias correction methods, we use the RMSE, which is the square root of the mean square error, defined as:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i)^2} \tag{1}$$

where x is some estimated quantity compared against a given target quantity y. Due to the square operator, the RMSE is a positive-valued function, which is less sensitive to low errors and penalizes large errors.

To compute a CI for the RMSE, we assume $x_i - y_i \sim \mathcal{N}\left(0, \sigma^2\right)$. If this holds true, the quantity $\frac{n\mathrm{RMSE}^2}{\sigma^2}$ follow a χ^2_n distribution with n degrees of freedom. From this, we get the CI $\left[\sqrt{\frac{n}{\chi_{1-\frac{\alpha}{2},n}}}\mathrm{RMSE}, \sqrt{\frac{n}{\chi^2_{\frac{\alpha}{2},n}}}\mathrm{RMSE}\right]$.

4.2 Models

4.2.1 Ordinary least square regression (OLS)

A linear regression model aims to predict the dependent variable (Y) from several independent variables (X_i) using a linear equation:

$$Y = a + b_1X_1 + b_2X_2 + ... + b_iX_i$$

, with the intercept a and the coefficients of the independent variables b_i as the parameters. The parameters are determined by minimizing the squared difference between the dependent variable and observations. It is assumed that the independent variables exert no co-linearity and have normally distributed homoscedastic errors. Note for clarity that the OLS was fit using the observed temperature values, not the residuals.

4.2.2 Neural Networks

Neural Networks (NNs) encompass a large set of models that maps input to output, with the criteria being that the mapping is composed of differentiable functions and that the computational graph is directed and acyclical, generally denoted

$$f(\mathbf{x};\theta) = f_L\left(f_{L-1}\left(\cdots\left(f_1\left(x\right)\right)\cdots\right)\right) \tag{2}$$

Where Equation. 2 represents a multilayer perceptron. A major benefit of fitting a NN to our dataset is the ability to make non-linear connections in the dataset.

We construct two NN models for correcting the data, a shallow and a deep expansive path model. Optimal models are determined based on the validation-loss. The neural networks are trained with the MSE loss, using the ADAM optimizer. If the learning rate reaches a plateau, we half the learning rate. Both NNs are constructed from blocks with each block defined as: fully connected layer, batch normalization, ReLU activation function, and dropout. We define the shallow NN as a one block model, the deep NN containing multiple blocks, and explore the hyperparameter space through a grid-search across learning-rates $(0.001, 0.01, 0.1)$, epochs $(20, 100, 200)$ and number of neurons in the hidden layer(s) $(4, 8, 16, 32, 64)$ or

([16, 32], [16, 32, 64], [16, 32, 64, 128], [32, 64, 128], [32, 64]). Note that for the deep NN only the expansive path is denoted, but the network is symmetrical.

We obtain the following hyperparameters for the two NN. The shallow model has 64 hidden units, is trained for 100 epochs with a learning rate of 0.01. The deep NN is trained for 200 epochs, a learning rate of 0.01 and five hidden layers containing $(32, 64, 128, 64, 32)$ nodes each.

4.2.3 eXtreme Gradient Boosting

Decision trees are a non-parametric approach to machine learning. eXtreme Gradient Boosting (XGBoost) is an open source implementation of gradient boosted trees (Chen & Guestrin, 2016), which in short sequentially fits additive models weighted by the previous models' misclassification. The resulting prediction is the sum of all models.

We conduct a grid-search across relevant parameters: num. estimators $(100, 500)$, max depth $(3, 6, 9, 12)$, learning rate $(0.01, 0.1, 0.2)$, subsample per tree $(0.8, 1.0)$ and fraction of features per tree $(0.6, 0.8, 1.0)$. Our optimal XGBoost model has 100 estimators, a learning rate of 0.1, depth of 6, 80% of the data subsampled, and all features used per tree.

5 Results and Discussion

5.1 Time series analysis

The time series of observed and forecast temperatures are shown in Fig. 1, with longer periods of observation for the training buoys 03 and 07 compared to the test buoy 10. The commonly recorded period shows similar temperature and SIC values for each buoy. However, recordings of the training buoys deviate thereafter, resulting in different statistics between training and test sets (mean, sd, Q25, Q75). In the test buoy, the forecast commonly overestimates the air temperatures, while the general statistics suggest a reasonable reconstruction of the temperature in the training buoys. Visual inspection suggests an overestimation of the forecast for cold temperatures, whereas warmer temperatures (ca. $0C$) are underestimated.

5.2 Bias correction of AROME Arctic

The histograms in Fig. 2 show the observed and forecast temperatures for the training and test data. For both distributions, the probability density around $0C$ of the forecast exceeds the observed values, whereas the overestimation is higher for the training than the test set. This also shows in the CDF plots, where the maximum difference in the training data is around $0C$, resulting from an overestimation by the forecast. For the test data, the maximum difference in the CDFs is around $-2.8C$, where the forecast underestimates the temperature. This is not intuitive from the histogram, but probably results from the slight bimodalality of the observations. The plots imply that the forecast underestimates negative temperatures, while it overestimates at temperatures around $0C$, confirming the visual inspection of Fig. 1.

To quantify the implied differences between forecast and observation, the KS-test determines whether they originate from the same distribution. As shown in 1, the KS-test rejects H_0 that the observations and forecasts belong to the same distribution with a CI of 95%.

Fig. 3 shows the time series of the residuals and the histogram, which imply a normal distribution and no bias for the training set. However, regarding the test set, a slight overestimation of the forecast is implied, especially by the negative skew in the histogram.

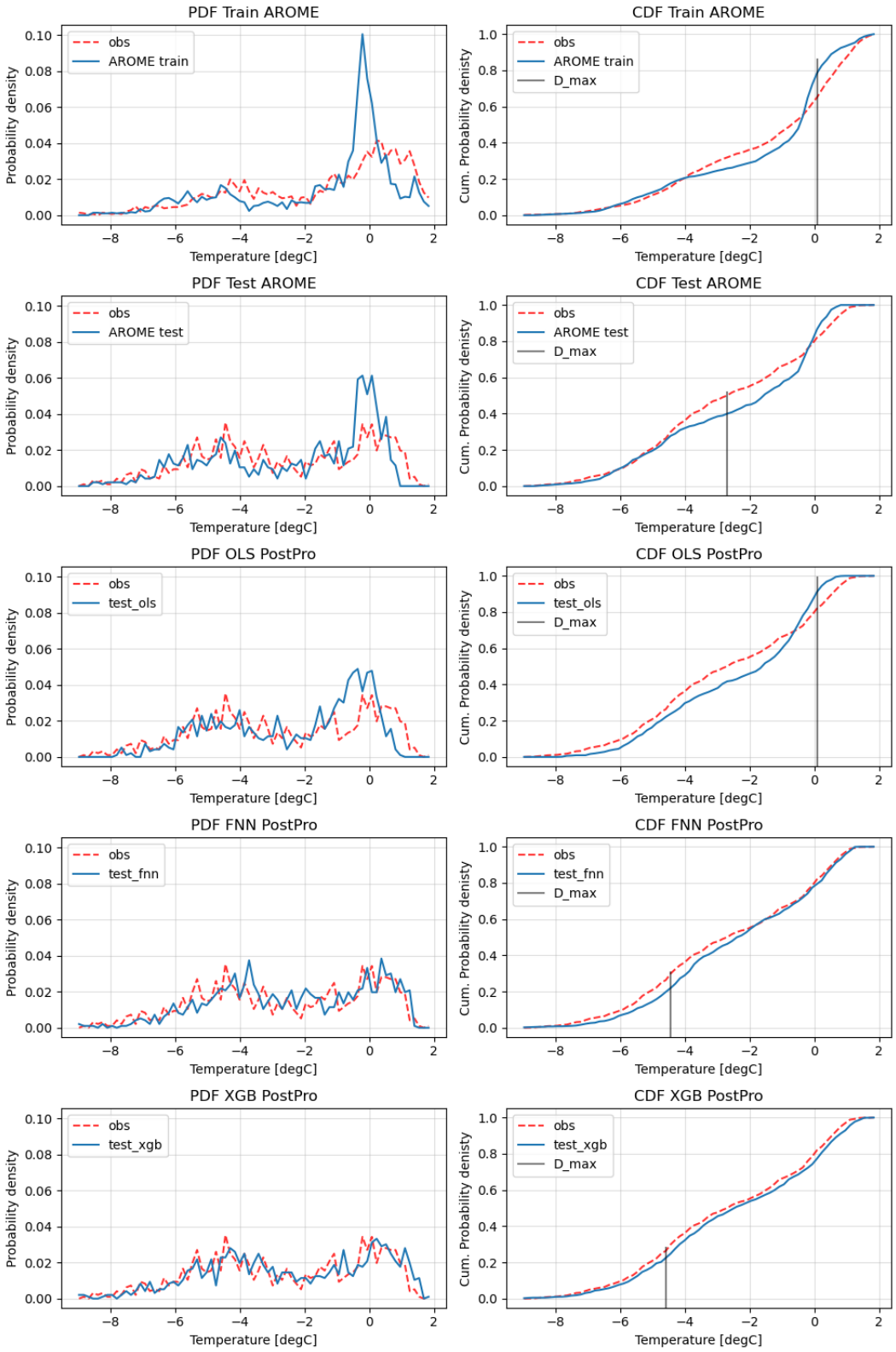

Figure 2. Histograms and cumulative probability density functions comparing the observed and forecast temperatures. The distributions are shown for the uncorrected forecast and observations of the training and test set, and for the bias-corrected forecast of the test data. Just the best performing model respective each correction method is shown here. The CDFs also indicate the maximum deviation between observed and forecast temperatures, as used in the Kolmogorov-Smirnov test.

Table 1. RMSE and KS-test results from the different temperature estimation. Processing regarding whether the forecast has been bias corrected or not, and with which method.

Correction	RMSE [95% $C.I.$]	KS Test stat. D_{max}	Rej. crit.: Crit. value $c_\alpha(n,m)$	$D_{max} > c(\alpha)\sqrt{\frac{n+m}{n*m}}$ $H_0 : F_1(x) = F_2(x)$
Unprocessed	1.203 [1.173 − 1.235]	0.135	0.035	Rejected
Unprocessed	1.271 [1.216 − 1.33]	0.109	0.062	Rejected
OLS corrected	1.194 [1.143 − 1.25]	0.105	0.062	Rejected
NN corrected	0.963 [0.962 − 1.008]	0.098	0.062	Rejected
FNN corrected	0.656 [0.628 − 0.687]	0.081	0.062	Rejected
XGB corrected	0.377 [0.361 − 0.395]	0.051	0.062	Not rejected
XGB day only	0.867 [0.830 − 0.908]	0.075	0.062	Rejected
XGB phys only	0.778 [0.745 − 0.815]	0.067	0.062	Rejected

This skew is absent in the training set due to the extended observation of more higher temperatures, which were underestimated by the forecast.

An underestimation of observed positive temperatures is indicated by the QQ-plots shown in Fig. 4, especially for the training dataset. The overestimation of the negative temperatures in the test set, which was implicit from Fig. 2, can be seen again by the bulging of the corresponding QQ-plot, thus deviating from the identity line.

The PP-plot for the training set shows a strong bulge above the identity line after a cumulative probability density of 0.6, which is consistent with the overestimation of higher temperatures. This indicates a more conservative forecast of temperatures.

The presented results indicate a bias in the forecast air temperature by AROME Arctic when compared to observations in the MIZ. Hereafter, the results of various methods of bias correction to improve the forecast are compared to the observations.

Depicted in Fig. 2 are the histograms and CDFs of the bias-corrected forecasts. Comparing the results to the uncorrected test set, the OLS bias correction appears to improve the overestimation regarding the predictions of negative temperatures. However, around $0C$, it still overestimates observations, and also the empirical and forecast CDF do not align well. The maximum deviation of the CDFs shifts to around $0C$, which implies that the OLS bias-correction improves the forecast of the bimodal observed temperatures.

However, the OLS model gives forecasts similar to the uncorrected forecast, whereas the bias-corrected forecasts from the NNs and the XGB models improve the predictions significantly in terms of lowering the RMSE. By inspecting Table 1 we see that the deep NN (denoted FNN) significantly outperforms the shallow NN. The test statistics of the KS-test (1) reject the hypothesis that the observations and the forecasts corrected by

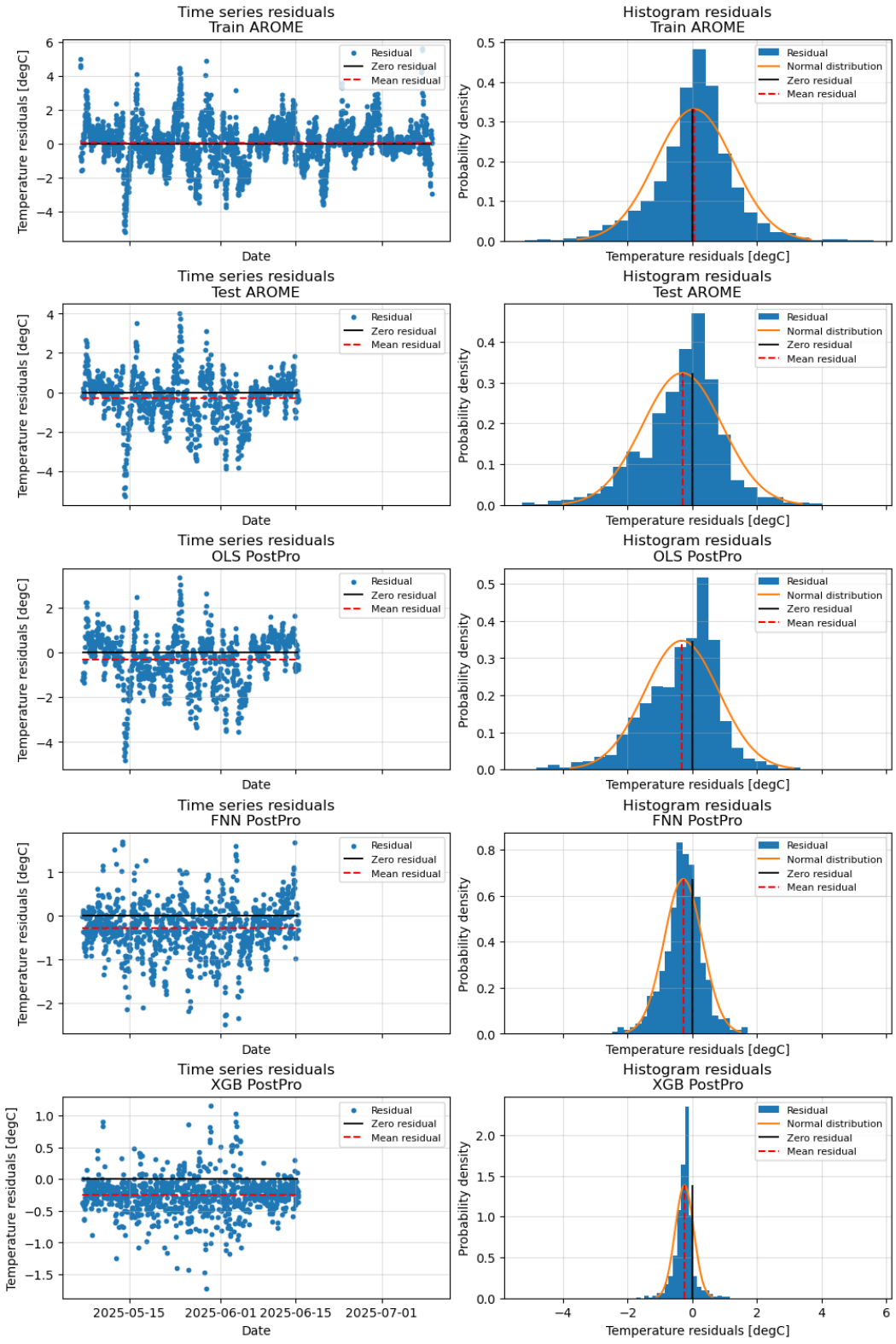

Figure 3. Time series of residual distribution and histogram of residuals training and testing data between the forecast and observed temperatures. The bias-corrected forecasts are shown for the test data, while just the best models of the respective methods are shown.

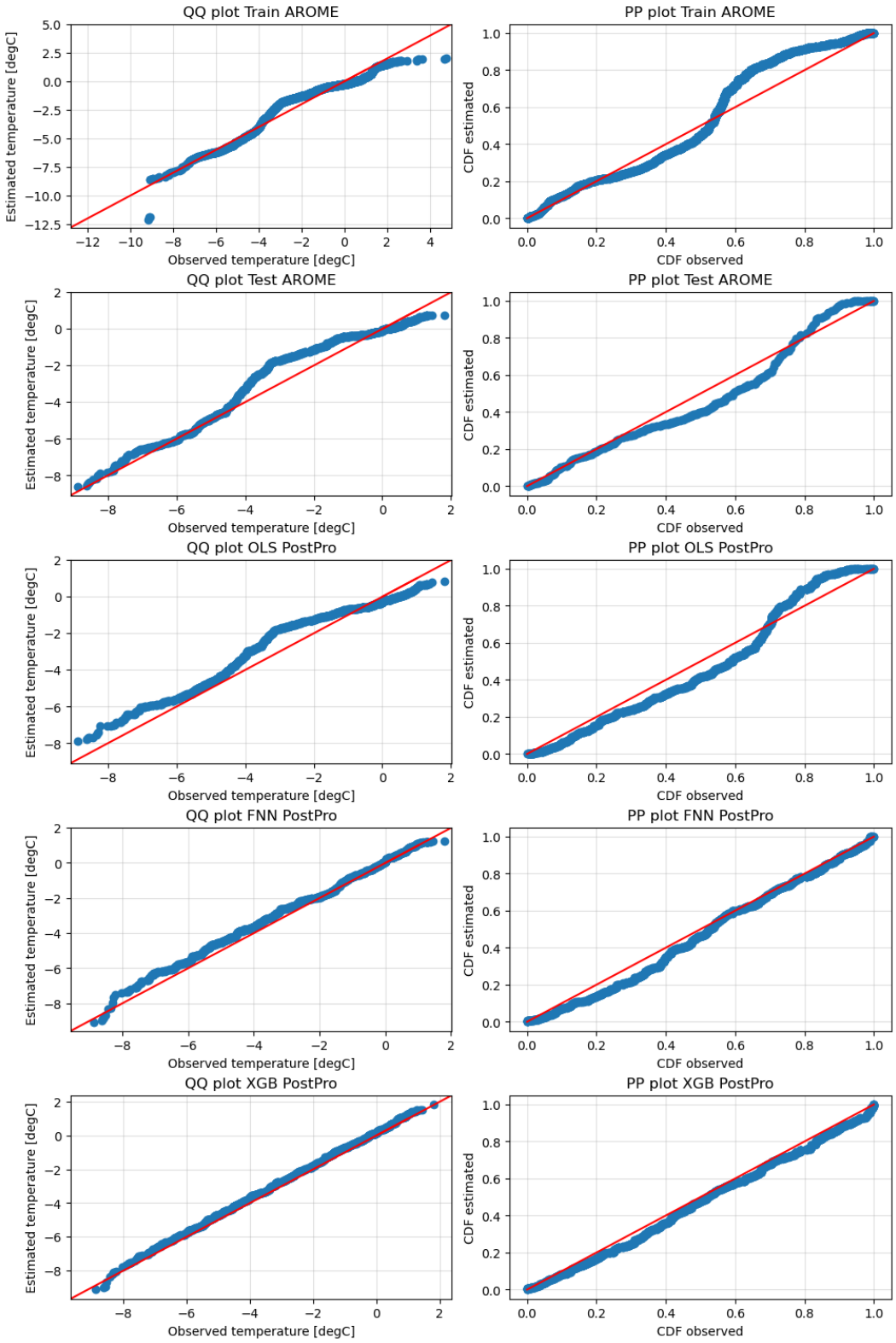

Figure 4. QQ- and PP-plots for the training and testing data respectively the observed and forecast air temperature from AROME Arctic. The bias-corrected forecast of the test data is shown against the observations, whereas just the best performing models of each correction method are shown.

the OLS or the NN belong to the same distribution as the observations. However, for the predictions from full XGB model, the KS-test indicates that the H_0 can't be rejected. Thus the following evaluation of the bias correction only considers the best performing model from the respective methods, i.e. FNN and XGB.

These models give histograms and CDFs similar to the observed temperatures. Minor differences can be seen in the negative temperature range (-4 to $-6C$) for the FNN predictions, while the positive temperature tail is reconstructed well. The CDF of the XGB model implies an improvement of the forecast in the negative temperature range, while it seems to generally underestimate the observations, since the CDF is always below the CDF of the observations.

The residuals of the OLS forecast are similar to the uncorrected forecasts comparing the times series and histograms of the residuals (y-axes not scaled). As for the uncorrected forecast, the negatively skewed residuals with a negative mean indicate an overestimation. The bias correction by the FNN results in residuals, which are more stochastically and narrowly distributed around the mean. Even though the residuals are normally distributed by visual inspection, the residual mean still indicates an overestimation of the corrected forecast. However, the residuals from the XGB model show an even narrower peak in the histogram and distribute over a closer temperature range throughout the evaluation period.

In term of the OLS correction, the QQ-plot resembles the uncorrected QQ-plot, except regarding the negative tail, where the corrected forecast overestimates the observed temperatures 4. Correspondingly, the probability of low temperature values is underestimated in the lower tail of the PP-plot, while it's generally similar to the uncorrected forecast. The decreased estimated probability mass in the lower half of the plot results in an increase of the probability mass in the upper tail. This can be interpreted as a stronger underestimation in the low temperature range and a stronger overestimation for higher temperatures.

For the FNN forecast, the QQ-plot follows the identity line closely, except for values in the lower temperature range, where an overestimation is implied. For higher temperatures, this overestimation does show in the PP-plots as well, where too little probability mass is present in the lower end of the plot, i.e., the corrected forecast overestimates low temperatures.

The QQ- and PP-plots of the XGB bias-corrected forecast closely follow the identity line throughout the temperature range. However, the QQ-plot lies slightly above the identity line, while the PP-plot is slightly below it. This implies that the corrected forecast temperatures generally underestimate the observed (QQ-plot), which results in slightly too little probability mass of the estimated values in the PP-plot. Even though this correction shows the slightest deviation, it still shows signs of being trained on a longer dataset, which experienced a period of higher temperatures that might have leaked into the predictions. This shows as well in the mean residuals 3, where all bias-corrected forecasts still overestimate the observed temperatures.

5.3 Model interpretation

Decision-tree based models are generally more interpretable than NNs. We utilize the "gain" metric in order to assess feature importance during training. Specifically, the gain parameter measures how much each predictor, on average, reduces the model-loss during training. We show the feature importance of the fitted XGBoost model in Fig. 5a).

Based on the importance scores in Fig. 5a), we fit two new XGBoost models from a predictor subset to distinguish the importance of physical and temporal predictors. Fig.

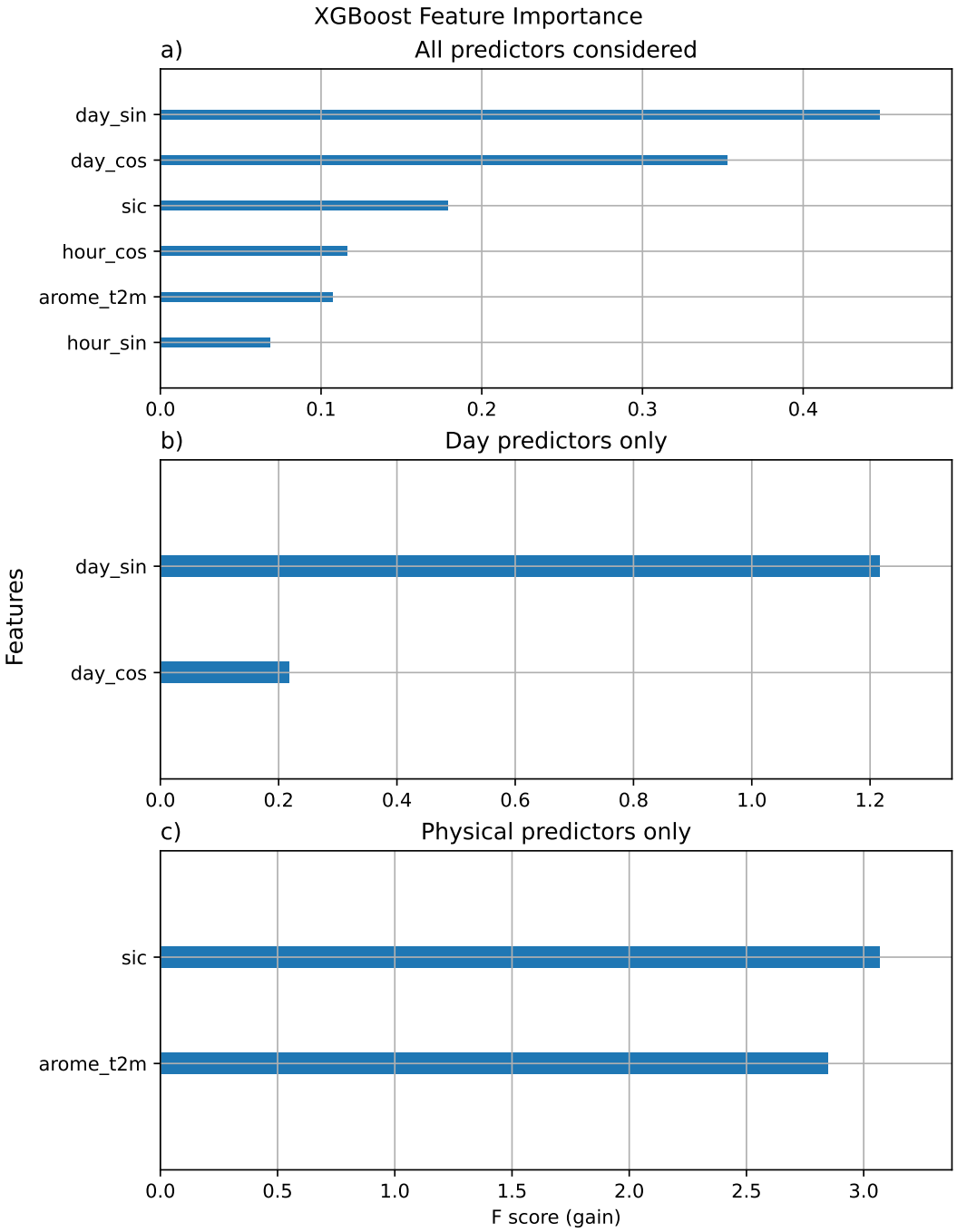

Figure 5. Feature importance for three XGBoost models, fitted with different combinations of predictors. Higher gain indicate greater contribution to loss reduction. Importance is measured through relative importance between predictors.

5b) and c) show the gain for doy and physical predictors only. We see in Fig. 5b) that the relative importance is highly skewed in favour of the sine component of the day, moreover for the physical parameters in Fig. 5c) both predictors are of relative equal importance. The ordering of predictor importance is unchanged between the sets.

Inspecting Table. 1, only the XGBoost model trained on all predictors achieves significant KS results, and that both XGBoost sub-models get higher RMSE. Comparatively, the model trained on physical predictors achieves significantly lower RMSE than the model trained on doy predictors. Thus simply fitting on the doy is not enough to construct a strong model, despite the structural similarity exerted by the time series in Fig. 1 suggesting otherwise. Hence, the XGBoost model requires physical conditions in order to make strong predictions, and we further speculate that this could increase the generalizability of our method to extend to conditions not currently captured by the considered buoys.

6 Conclusions

The observed temperature values are significantly different from modeled values in the MIZ. Different statistical models for bias correcting temperature forecasts have been developed. Most models significantly improved the RMSE beyond the uncorrected data, yet only for the XGBoost, the KS test suggested that the corrected data might have been drawn from the same distribution as the observations.

Due to the non-quality control of the dataset, the implications of our results are limited due to the structural similarity between the used buoys as they were deployed spatially and temporally adjacent. We urge future research to expand upon the dataset, including different geographical and met-ocean regimes. Nevertheless, we demonstrated our novel approach to improve forecast quality in the MIZ using data driven methods.

7 Resources

The developed code can be accessed at the following GitHub repository `https://github.com/AreFrode/GEO9300_project`.

References

Batrak, Y., & Müller, M. (2019, September). On the warm bias in atmospheric reanalyses induced by the missing snow over arctic sea-ice. *Nature Communications*, *10*(1). doi: 10.1038/s41467-019-11975-3

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (p. 785–794). ACM. Retrieved from `http://dx.doi.org/10.1145/2939672.2939785` doi: 10.1145/2939672.2939785

Hieta, L., & Partio, M. (2025). Operational machine learning post-processing of short-range temperature, humidity, wind speed and gust forecasts. *Meteorological Applications*, *32*(4), e70074. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.70074` (e70074 MET-24-0162.R1) doi: https://doi.org/10.1002/met.70074

Müller, M., Rabault, J., Palerme, C., & Tjernström, J. (2025, September). *Svalmiz-25 svalbard marginal ice zone campaign 2025 – cruise report.* arXiv. doi: 10.48550/ARXIV.2509.10016

Müller, M., Batrak, Y., Kristiansen, J., Køltzow, M. A. Ø., Noer, G., & Korosov, A. (2017). Characteristics of a convective-scale weather forecasting system for the european arctic. *Monthly Weather Review*, *145*(12), 4771 - 4787. Retrieved from `https://journals.ametsoc.org/view/journals/mwre/145/12/`

mwr-d-17-0194.1.xml doi: 10.1175/MWR-D-17-0194.1

Müller, M., Rabault, J., Palerme, C., & Tjernström, J. (2025, September). *Dataset: Svalmiz-25 svalbard marginal ice zone campaign 2025 - a distributed network of temperature, waves in ice, and sea ice drift observations.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.17087019 doi: 10.5281/zenodo.17087019

Palerme, C., & Müller, M. (2021). Calibration of sea ice drift forecasts using random forest algorithms. *The Cryosphere*, *15*(8), 3989–4004. Retrieved from https://tc.copernicus.org/articles/15/3989/2021/ doi: 10.5194/tc-15-3989-2021

Schyberg, H., Yang, X., Køltzow, M., Amstrup, B., Bakketun, Å., Bazile, E., ... Wang, Z. (2020). *Arctic regional reanalysis on single levels from 1991 to present [dataset].* Copernicus Climate Change Service (C3S) Climate Data Store (CDS). Retrieved from https://doi.org/10.24381/cds.713858f6 doi: 10.24381/cds.713858f6

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., ... Ylhaisi, J. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, *102*(3), E681 - E699. Retrieved from https://journals.ametsoc.org/view/journals/bams/102/3/BAMS-D-19-0308.1.xml doi: 10.1175/BAMS-D-19-0308.1