UNIVERSITÄT LEIPZIG

Weather Forecast Post-Processing

Group 8

Deepak Reddy Chinthaparthy, Khuzaima Aziz, Ali Bagheri ,Muhammad Abdul Wasay Khalid

Task 1*: Describe the data and prediction change.*
Question 1.a*: What is the dataset and the prediction problem about scientifically?*
The EUPPBench dataset is designed to facilitate the benchmarking of various post processing techniques used in weather forecasting. It is also used to address the prediction problem of improving the accuracy of deterministic weather forecasts. In our project we applied PCR and Boosted Trees to correct errors and biases in raw numerical weather prediction (NWP) model output . We used 2m air temperature as a predictor variable and enhanced the accuracy of modeled data and the final goal is to improve accuracy of forecasting related to air temperature.

Question 1.b: *What's the data source and what does that tell you about the type of the data?*
This dataset includes forecast data (modeled) from the Integrated Forecasting System (IFS) of the ECMWF and also observational data from the German Weather Service (DWD). The dataset includes both predicted model data and real-world observation data. This combination allows for effective comparison and refinement of weather forecasts, essential for improving the accuracy of weather prediction models through post processing techniques.

Task 2: *Describe the background of your assigned linear and non-linear methods.*
Task 2.1: *For the linear method.*
Question 2.1.a: *What is the key idea (that is, theoretical basis) behind your method?*
Our linear method, Principal Component Regression (PCR), improves the regression model by transforming the original predictor variables into a new set of uncorrelated variables (principal components) and then performing regression on these components.

Question 2.1.b: *In which circumstances is it advisable to apply your method?*

- If we have many predictors- reduces dimensionality and captures most of the variance
- If we have correlated predictors
- when the relationships between predictors and the response variable are complex

Question 2.1.c: *In which cases may your method fail?*

- If the principal components do not explain a significant portion of the variance in the predictor variables, the model built on these components may not capture the underlying relationships well.
- PCR ignoring components that explain less variance but are still important for the response variable.
- Principal components are linear combinations of the original variables, making it difficult to interpret the contribution of each original predictor to the response variable.
- If the true relationship is nonlinear, PCR may fail to capture it effectively.

Question 2.1.d: *How does your method guard against overfitting?*
By selecting only the top principal components that explain the most variance. By using uncorrelated components, PCR stabilizes the regression coefficients. In our method , using 10 fold cross-validation to determine the number of principal components helps ensure that the selected components generalize well to unseen data, further guarding against overfitting. PCR provides a bias that reduces variance, effectively balancing the trade-off and guarding against overfitting.

Question 2.1.e: *Which are the tuning parameters of your method?*
The number of principal components (k) is chosen based on cross-validation, aiming to balance the bias-variance trade-off. Here we have 14 predictors and try to find the optimal number of principal components.

Task 2.2: *For the non-parametric method.*
Question 2.2.a: *What is the key idea (that is, theoretical basis) behind your method?*
Boosted Trees is a type of ensemble learning which uses the classification and regression trees (CART) method. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models is added.

Question 2.2.b: *In which circumstances is it advisable to apply your method?*

- When the dataset is complex with nonlinear relationships.
- When the dataset includes both numerical and categorical features.
- Applications requiring real-time predictions, the prediction phase is relatively fast compared to training.

Question 2.2.c: *In which cases may your method fail?*

- Dealing with large amounts of data it can be computationally expensive and time consuming.
- Complexity of the models makes them difficult to interpret.

Question 2.2.d: *How does your method guard against overfitting?*

- Shrinkage (Alpha): The impact of each individual tree is reduced by multiplying its contribution by a small learning rate,
- Number of trees (B): Using more trees enhances model performance up to a point, but using too many trees can result in overfitting.
- Maximum Depth (d) : Limiting each tree's maximum depth stops the model from capturing excessive noise.

Question 2.2.e: *Which are the tuning parameters of your method?*

- n.trees: Integer specifying the total number of trees to fit.
- Interaction.depth: Integer specifying the maximum depth of each tree.
- Shrinkage: known as the learning rate or step-size reduction
- bag.fraction: the fraction of the training set observations randomly selected to propose the next tree in the expAnsion. This introduces randomnesses into the model fit.
- cv.folds: Number of cross-validation folds to prefer.
- n.minobsinnode Integer specifying the minimum number of observations in the terminal nodes of the trees.

Task 3: *Describe the process of setting up and tuning your statistical learning model.*
Task 3.1: *Explore your materials.*
Question 3.1.a: *Which data-processing steps have been included in the basic setup provided for your dataset?*
The data include forecast and observation variables, which were reshaped and recombined in a tabular test and train dataset format and to capture the seasonality, sin and cos variables were added.

Question 3.2.a: *How did you set up the predictor variables for the model? Suggest at least two different setups/combinations of explanatory variables/features that you consider. Argue for the specific choices.*
PCR is not suitable for feature selection. Therefore, we explored two different setups: the first setup included all 14 variables (12 predictor variables from the model data and 2 additional predictor variables for seasonality), while the second setup included only the top five highly correlated variables: "dmo", "cos.doy", "strd6", "mx2t6", and "mn2t6"

Question 3.2.b: *Which hyperparameter(s) do you optimize?*
Choosing the optimal number of principal components (k)

Question 3.2.c: *Describe your tuning procedure: How was data split (CV, train/test-split), and why did you choose this splitting? Consider the special characteristics of your data set.*
We are provided with test data and train data . Training data starts from 1998 and ends in 2016.Testing data is for only one year 2017.Ratio is 5.2 %. I.e (one year / 19 years)

Question 3.2.d: *Which setup did perform best? Of this setup, what were the best performing hyperparameter values?*
Setup 1 worked well i.e with all 14 variables used as predictors. We have only one hyperparameter k, no of principal components values ranging from 10 to 13 for all 49 stations.

Question 3.2.e: *Briefly discuss possible reasons for the good performance of the best setup.*
Although adding more predictors in linear regression typically reduces training error and risks overfitting, in our case, we manage overfitting with tuning hyperparameter (k). This approach allowed us to successfully use all the variables which increased the performance

Task 3.3: *Describe how your non-parametric model was set up.*
Question 3.3.a: *How did you set up the predictor variables for the model? Suggest at least two different setups/combinations of explanatory variables/features that you consider. Argue for the specific choices.*
Setup 1 involved identifying highly correlated variables with observed values from 49 weather stations, including dmo (2m temperature), mx2t6 (maximum 2m temperature), mn2t6 (minimum 2m temperature), strd6 (surface thermal radiation), and cos.doy (cosine of day of year). We standardized hyperparameters with 300 trees, interaction depth of 3, shrinkage of 0.1, and bagging fraction of 0.7. Each station adjusted the number of trees using gbm.per, and keeping all of hyperparameter constant.
Setup 2 Here, we conducted feature selection through cross-validation using GBM as the model and selected the most important features (i.e., top 5). We then tuned parameters for each station, which differed from setup 1 where all hyperparameters were the same except for the number of trees. This setup took considerably more time (~2 hours) compared to the first setup which took 10 minutes.

Question 3.3.b: *Which hyperparameter(s) do you optimize?*

- N.trees
- Interaction.depth
- shrinkage
- n.minobsinnode

Question 3.3.c: *Description of the tuning procedure: How was data split (CV, train/test-split), and why did you choose this splitting? Consider the special characteristics of your data set from atmospheric and climate research*.

Training data starts from 1998 and ends in 2016.Testing data is for only one year 2017.Ratio is 5.2 %. I.e (one year/19 years)

Question 3.3.d: *Which setup did perform best? Of this setup, what where the best performing hyperparameter values?*

The results were astonishingly similar. Nevertheless, we submitted the predictions of setup 2 even though we obtained the almost same results with setup 1 because we performed extensive feature selection and hyperparameter tuning on each station, expecting it to perform well on untested data.
The hyperparameter for each station were different because we tuned individually for each station Example of one station id 460 used parametric grid search for other parameters. Typical Grid Parameter Search Used in Our Model Uses this range of values depicted in bracket and finding out the optimal parameter:
n.trees = 300 (100, 200, 300)
interaction.depth = 4 (3, 4, 5)
shrinkage = 0.1 (0.01, 0.1)
minobsinnode = 10 (5, 10)

Question 3.3.e: *Briefly discuss possible reasons for the good performance of the best setup.*

Result in both the setup were similar , possible reason for comparative performance is locally tuning hyper parameters and selecting features of high importance in each station.

Task 4: *Describe and discuss your results*

Question 4.1: *Note down the mean squared error (MSE) for your best performing linear and nonparametric model.*

The mean square error (MSE) for the linear model PCR is "4.80" and for the non-parametric model GBM the MSE is “5.20”

Question 4.2: *Discuss the performance difference of the linear and non-parametric method. Does the performance difference meet your expectations?*

Linear method PCR is comparatively simple, rigid , faster and performed better whereas non-parametric method GBM is complex, flexible and time consuming. We expected GBM to work better as it can address the nonlinear nature of meteorological data but it doesn't in our case.

Question 4.3: *Is one of the assigned methods an appropriate choice for the dataset and question at hand? Why, or why not, based on shortcomings or advantages of the linear and non-parametric method? Suggest alternatives to address the shortcomings of your method.*
The PCR method cannot be used for feature selection so if we use lasso regression as feature selection and perform PCR later may lead to better results. For the non-parametric all methods are computationally expensive and it would have the same results.

Question 4.4: *Explain how the different features in one of your models contribute to the prediction provide at least one statistical/data-driven argument, and one argument based on your Earth system science understanding.*

Feature	Variable name	no of stations	Correlation coefficient(R2)
2 meter air temperature	dmo	49	0.91
2 meter air minimum temperature	mn2t6	49	0.91
2 meter air maximum temperature(mx2t6)	mx2t6	49	0.92
Surface net thermal radiation	str6	32	0.05
Surface latent heat flux	slhf6	18	-0.311

dmo, mn2t6, and mx2t6 are used as features in the model across all 49 stations. These variables represent 2-meter air temperature measurements—specifically, mean, minimum, and maximum temperatures, respectively. They are statistically correlated and essential in modeling temperature dynamics, including their diurnal variations.Str6 is included as a feature in the model for 32 stations. Surface net thermal radiation (str6) plays a crucial role in the energy budget of the Earth's surface and is closely related to temperature changes. Similarly, slhf6 is selected as a feature in the model for 18 stations. Surface latent heat flux (slhf6) is vital for understanding energy exchanges, particularly latent heat release associated with phase changes in water, which influence temperature dynamics.

Question 4.5: *Assume that you would take your trained statistical model, and would apply it to actual Earth system science observations (or forecasts made by a different weather model in Group Project4). How would you expect your model to perform? What could be done to further improve your model's performance?*
Our model is expected to capture general trends in variability, but its accuracy may be limited due to biases from other weather models and the absence of additional parameters used in those models. To improve performance, incorporating data from other models for fine-tuning and introducing noise to simulate biases could be beneficial.

Task 5: *Apply the model to new data*
Question 5.1: *Use your linear and non-parametric model to obtain a prediction for the „test" dataset, or which no target values („Y's") are provided. Submit your predictions.*
We submitted our result as my.PCR.MOS_group_8.rds for linear and my.GBM..MOS_group_8.rds for non-parametric model.

Question 5.2: *Please give an estimate of the mean squared error (MSE) you are expecting your model to achieve for the prediction of the unseen test data. Justify your estimate briefly.*

We assume that the MSE for the unseen test data will be similar to the MSE estimated for the year 2017, as the time period is consistent. For 2017, the range of MSE for PCR is 3.27 to 9.85, and for GBM, it is 3.35 to 10.90. The total MSE for all stations is 4.80 for PCR and 5.20 for GBM. Therefore, we expect the MSE for the unseen test data i.e for 2018 to fall within these ranges. The range is between 3.27 to 10.90, with the total MSE between 4.80 and 5.20.