CONVOLUTIONAL AUTOENCODERS FOR SPATIALLY-INFORMED ENSEMBLE POST-PROCESSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensemble weather predictions typically show systematic errors that have to be corrected via post-processing. Even state-of-the-art post-processing methods based on neural networks often solely rely on location-specific predictors that require an interpolation of the physical weather model's spatial forecast fields to the target locations. However, potentially useful predictability information contained in large-scale spatial structures within the input fields is potentially lost in this interpolation step. Therefore, we propose the use of convolutional autoencoders to learn compact representations of spatial input fields which can then be used to augment location-specific information as additional inputs to post-processing models. The benefits of including this spatial information is demonstrated in a case study of 2-m temperature forecasts at surface stations in Germany.

1 INTRODUCTION AND MOTIVATION

Most weather forecasts today are based on ensemble simulations from numerical weather prediction (NWP) models, consisting of a set of deterministic forecasts that differ in initial conditions or model physics. Despite continued improvements (Bauer et al., 2015), ensemble predictions continue to exhibit systematic errors such as biases or a lack of calibration. The process of correcting such errors to obtain more accurate and reliable forecasts is referred to as post-processing (Vannitsem et al., 2018). Post-processing models use ensemble predictions from the NWP system as inputs, and produce a probability distribution as their output. Post-processing has become a standard practice in research as well as operations, and is an integral part of weather forecasting today. Parametric approaches from statistics where the forecast distribution takes the form of a probability distribution with parameters depending on summary statistics of the ensemble predictions of the target variable have been developed for a large variety of weather quantities. Over the past years, much work has been spent on flexible machine learning techniques for post-processing which enable the incorporation of additional predictor variables beyond ensemble forecasts of the target variable, and have demonstrated superior forecast performance (Haupt et al., 2021; Vannitsem et al., 2021). Much recent research interest has been focused on neural network (NN)-based distributional regression approaches first proposed in Rasp & Lerch (2018), where NNs learn nonlinear relationships between arbitrary predictor variables and forecast distribution parameters in a data-driven way.

All of these post-processing methods share a common limitation: To provide predictions at individual locations (typically weather stations or grid points), they require localized ensemble forecasts, which are obtained by interpolating the NWP ensemble members' two-dimensional forecast fields to the target locations. However, the large-scale spatial structure and predictability information[1] present in the physically consistent forecast fields from the ensemble simulations are lost in this interpolation step. We propose the use of convolutional autoencoders to learn low-dimensional latent representations of the spatial forecast fields. The learned representations are then used as additional predictors to augment a NN-based post-processing model with information about the spatial structure of relevant forecast fields. The proposed model architecture is applied in a case study of 2-m temperature forecasts at surface stations in Germany, and compared to state-of-the-art post-processing models without spatial inputs.

The remainder of the paper is organized as follows. Section 2 introduces the data. In Section 3, we describe the autoencoders and their combination with post-processing models. The main

[1] e.g., flow-dependent error characteristics and weather regimes (Rodwell et al., 2018; Allen et al., 2021)

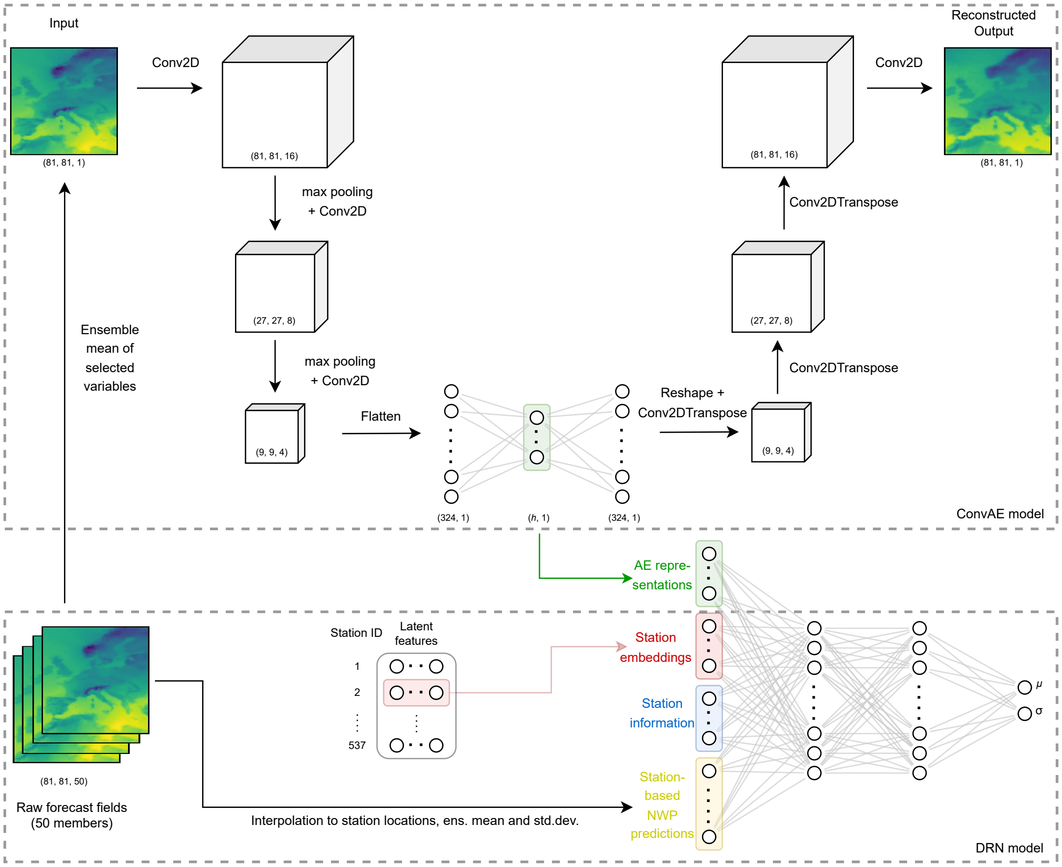


Figure 1: Schematic illustration of the DRN+ConvAE model.

results are presented in Section 4, and Section 5 concludes with a discussion. Python code with implementations of all methods is available online (link removed to preserve anonymity).

2 Data

We use the dataset from Rasp & Lerch (2018) focusing on 2-m temperature (T2M) forecasts at surface stations in Germany at a forecast lead time of 48 h. Forecasts from the global European Centre for Medium-Range Weather Forecasts 50-member ensemble initialized at 00 UTC every day form the basis for two types of predictor variables. To obtain location-specific predictors, following Rasp & Lerch (2018), we interpolate ensemble forecast of 17 meteorological variables to the observation station locations, see their Table 1 for an overview of the available variables. In addition, we also use the spatial forecast fields of T2M, geopotential height at 500 hPa (Z500), and the U- and V-wind at 850 hPa (U850 and V850) as a second dataset of spatial inputs. Those variables were chosen broadly based on meteorological intuition, and are available on $0.5^\circ \times 0.5^\circ$ grid from -10E to 30E and from 30N to 70N, which roughly covers Europe and parts of the surroundings.

Observation data of T2M for 537 weather stations in Germany are used to evaluate the forecasts. Information about the station coordinates, altitudes and orography (altitude of the model grid point) are derived as additional input predictors for the post-processing models. With ensemble predictions available from 3 January 2007 to 31 December 2016, we follow the setup in Rasp & Lerch (2018) and use data from 2007–2015 as training dataset, and data from 2016 as test dataset.

3 Methods

We focus on post-processing methods within the parametric distributional regression framework proposed by Gneiting et al. (2005). In their ensemble model output statistics (EMOS) approach, the conditional distribution of the variable of interest y, given ensemble predictions $\boldsymbol{X}$, is modeled by a parametric distribution, $F_{\boldsymbol{\theta}}$, with parameters $\boldsymbol{\theta} = g(\boldsymbol{X})$ depending on the ensemble predictions via a link function g. The standard EMOS model for temperature utilizes ensemble forecast of temperature $\boldsymbol{X}^{\text{t2m}}$, as sole predictors and assumes a Gaussian forecast distribution $y|\boldsymbol{X}^{\text{t2m}} \sim \mathcal{N}(\mu, \sigma)$, the parameters of which are linked to the ensemble mean and standard deviation via affine functions $\mu = a + b \cdot \text{mean}(\boldsymbol{X}^{\text{t2m}})$ and $\sigma = c + d \cdot \text{sd}(\boldsymbol{X}^{\text{t2m}})$. The model coefficients a, b, c, d vary over stations (for local adaptivity), and are estimated by minimizing the mean continuous ranked probability score, $\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(z) - \mathbb{1}(y \leq z)\right)^2 \mathrm{d}z$, over a training set (Jordan et al., 2019).

3.1 Neural network methods for post-processing

A key limitation of the EMOS approach is that incorporating additional predictors beyond forecasts of the target variable is challenging since it would be necessary to specify the exact functional form of the dependencies of the distribution parameters on all input predictors. To address this limitation, Rasp & Lerch (2018) propose to obtain the distribution parameters as the output of a NN that is able to flexibly learn nonlinear relations between arbitrary input predictors and the distribution parameters in an automated, data-driven manner. Their distributional regression network (DRN) illustrated in the bottom part of Figure 1 is estimated as a single model jointly for all stations, using the CRPS as a custom loss function (D'Isanto & Polsterer, 2018). Station embeddings which map the station identifiers to a vector of latent features used as additional inputs to the NN generate local adaptivity in the jointly estimated model. The results presented in Rasp & Lerch (2018) and subsequent research demonstrate the improvements over state-of-the-art-approaches.

3.2 Convolutional autoencoders as information compressor

An autoencoder (AE) is a NN designed to learn a representation for a dataset by training the network to attempt to copy its input to its output. Internally, a hidden layer describes an h-dimensional encoding used to represent the input. Along with the encoder function, a decoder is learned that produces a reconstruction of the input data from the hidden layer. Here, we consider separate AE models for selected weather variables (T2M, Z500, 850U, 850V) which use the spatial fields of the ensemble mean forecasts as inputs. To account for the spatial structure of the input fields (with a size of 81×81 grid points), we use 2D-convolutional layers for the encoder and corresponding transposed convolutions for the decoder, and refer to the full model as convolutional autoencoder (ConvAE), illustrated in the top part of Figure 1. The ConvAE model is estimated separately in a first step, using the grid point-wise mean squared error as loss function. Min-max normalization is applied to the individual input fields in order to guide the ConvAE models to focus on the variability across space within the forecast fields, since the relevant information on the magnitude of the predicted values is present in the station-specific, interpolated predictors for the DRN model. Details on the model architecture and training are provided in the supplemental material. As a reference dimensionality reduction method, we implement a principal component analysis (PCA) approach, which is widely used for different applications in the atmospheric sciences (e.g., Wilks, 2011).

3.3 Incorporating spatial inputs into NN-based post-processing

To incorporate spatial information into the DRN model, mean forecast fields from the ensemble are used as input to the ConvAE model which was separately estimated in a first step and yields a corresponding latent space representation as output. This latent space representation is then used as additional input to the DRN model in addition to the station-specific (interpolated) predictions, the station information and the embeddings. This combined model will be referred to as DRN+ConvAE model and is illustrated in the entirety of Figure 1. We only consider spatial inputs from single predictors (T2M, Z500, 850U, 850V), T2M combined with Z500, or all of them. To ensure comparability, the architecture and training procedure for the DRN+ConvAE models are identical to those of the DRN model without spatial inputs. We proceed analogously for the DRN+PCA models. See Appendix A.1 in the supplementary material for details on the model architecture and estimation.

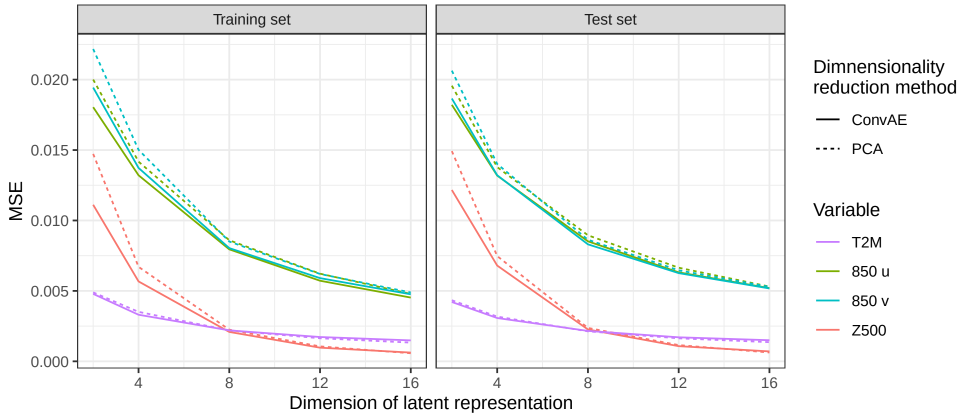


Figure 2: Grid point-wise MSE of reconstructions by the ConvAE and PCA models as function of the dimension h of the latent representation for several targets on the training set (left) and the test set (right).

4 Results

Figure 2 shows the mean squared error (MSE) of the ConvAE and PCA reconstructions. Since they are computed on the 0-1 scale of the normalized inputs, the errors of the different target variables are directly comparable, Not surprisingly, larger values of the dimension of the latent representation h yield better reconstructions. The T2M and Z500 forecast fields are generally easier to reconstruct from the low-dimensional representation than the wind fields which are characterized by small-scale structures. Overall, the ConvAE models show lower reconstruction errors compared to the PCA models. The differences are most pronounced for lower-dimensional representations, and forecast fields of Z500 and the wind components, but become negligible for larger values of h. Exemplary reconstructions of the ConvAE models are shown in the supplemental material in Appendix A.2. Compared to the corresponding PCA reconstructions (not shown), in particular small-scale variability is notably better represented by the ConvAE models for lower values of h.

Since our focus is on incorporating spatial features, we refer to Rasp & Lerch (2018) for detailed results on the DRN model, including comparisons to other post-processing approaches. Figure 3a summarizes our key results by showing the mean CRPS (averaged over all stations and dates in the test set, with lower values indicating better forecasts) of the DRN+ConvAE and DRN+PCA models as functions of the corresponding dimension of the latent representations for different sets of spatial inputs. Improvements over DRN can be observed for DRN+ConvAE models that include spatial inputs from T2M, Z500, and their combination, for latent dimensions $h \leq 8$. Generally, the forecast performance decreases with increasing h, likely since the most relevant information is already contained in lower-dimensional spatial representations. Note that the reconstruction quality of the ConvAE and PCA models might only be of minor importance for the post-processing task of the combined model, since the added reconstruction quality may be counteracted by making it more difficult for the DRN part of the model to extract the relevant information. Incorporating representations from wind fields notably deteriorates the results compared to the plain DRN model. Directly comparing DRN+ConvAE and DRN+PCA models, it is evident that in contrast to their ConvAE counterparts, adding PCA representations only provides some minor improvements over DRN for T2M and $h = 2$.

Focusing on a comparison of the local effects of incorporating spatial inputs, Figure 3b shows the station-specific improvement in the mean CRPS. We here compare one DRN+ConvAE model to DRN, a CRPS skill score (CRPSS) value of 0.1 thus indicates an improvement of 10% in the mean CRPS over DRN at that station. Including spatial inputs in the DRN+ConvAE model results in improvements at 96% of the stations, with Diebold-Mariano tests of equal predictive performance (Diebold & Mariano, 1995) indicating that around two thirds of those improvements are statistically significant at a level of 0.05. A comparison of the importance of the input features of the two models suggests that the DRN+ConvAE model is indeed able to extract some useful information from the ConvAE representations, see the supplemental material in Appendix A.3 for details.

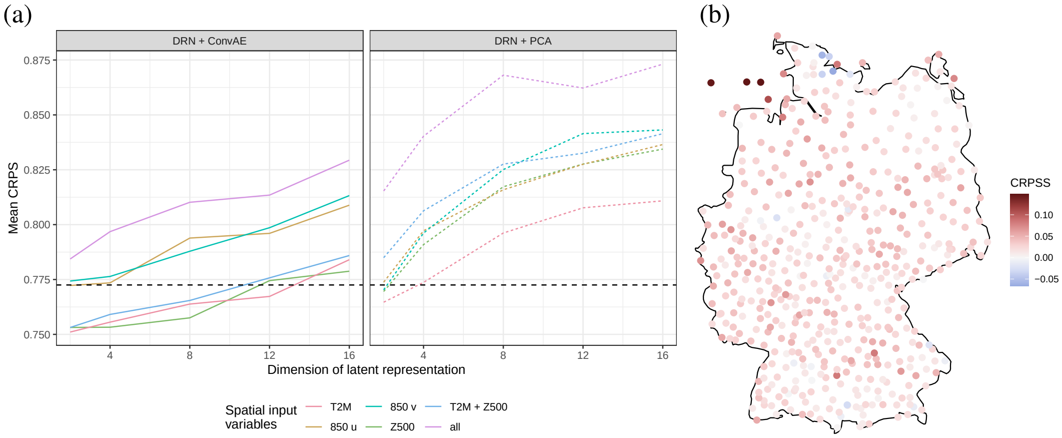


Figure 3: Mean CRPS of DRN+ConvAE (left) and DRN+PCA models (right) for different spatial inputs and values of h, with the CRPS of the DRN model shown as dashed black line (a); and map of station-specific improvements in terms of the CRPS of the DRN+ConvAE model with spatial T2M inputs for $h = 2$, using the DRN model without spatial inputs as reference (b).

5 Conclusions

We demonstrated how information from large-scale spatial forecast fields of meteorological variables can be incorporated into post-processing models via convolutional autoencoders as information compressors. Our post-processing models with added spatial inputs outperform state-of-the-art models that utilize station-specific predictors only. Regarding the design of the combined post-processing model with spatial inputs, we noted the critical need to balance the dimension of the latent representation of the spatial forecast fields: While larger values of h result in better reconstructions by the ConvAE models, they decrease the forecast performance of the corresponding post-processing model. In addition, increasing the embedding dimension compared to the model without spatial inputs was critical to obtain improvements, likely because this allows the post-processing models to learn make more locally adaptive use of the added spatial inputs. While our focus was on using NN-based post-processing models, it would also be interesting to incorporate the spatial representations as additional predictors into other post-processing models, such as quantile regression forests (Taillardat et al., 2016) or gradient boosting extensions of EMOS (Messner et al., 2017).

An alternative route towards post-processing models based on spatial input data is given by the direct use of convolutional NNs (CNNs). Forecast fields of several variables can be considered as input, and could be combined with a DRN part (variants of this approach have been proposed recently in Scheuerer et al., 2020; Veldkamp et al., 2021; Chapman et al., 2022; Li et al., 2022). A comparison to the methods proposed here is not straightforward since those approaches typically use gridded observations as target variables, but provides an interesting starting point for future research.

Finally, the ConvAE models proposed here raise several interesting methodological questions. While we only used mean fields as inputs, it would be interesting to consider probabilistic encoders in order to make better use of the ensemble structure of the ensemble members' forecast fields, which can be interpreted as samples from a spatial probability distribution. Further, a meteorological analysis of the learned representations of the spatial forecast fields, for example considering links to weather regimes, would not only be interesting from a meteorological perspective, but might also allow for better incorporating physical information and constraints into the forecasting models.

Acknowledgments

Acknowledgments removed to preserve anonymity

References

Sam Allen, Gavin R. Evans, Piers Buchanan, and Frank Kwasniok. Incorporating the North Atlantic Oscillation into the post-processing of MOGREPS-G wind speed forecasts. *Quarterly Journal of the Royal Meteorological Society*, 147(735):1403–1418, 2021.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

John Bjørnar Bremnes. Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Monthly Weather Review*, 148(1):403–414, 2020.

William E. Chapman, Luca Delle Monache, Stefano Alessandrini, Aneesh C. Subramanian, F. Martin Ralph, Shang-Ping Xie, Sebastian Lerch, and Negin Hayatbini. Probabilistic predictions from deterministic atmospheric river forecasts with deep learning. *Monthly Weather Review*, 150(1): 215–234, 2022.

Francis X. Diebold and Robert S. Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–263, 1995.

Antonio D'Isanto and Kai L. Polsterer. Photometric redshift estimation via deep learning-generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111, 2018.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

Sue Ellen Haupt, William Chapman, Samantha V. Adams, Charlie Kirkwood, J. Scott Hosking, Niall H. Robinson, Sebastian Lerch, and Aneesh C. Subramanian. Towards implementing artificial intelligence post-processing in weather and climate: Proposed actions from the Oxford 2019 workshop. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200091, 2021.

Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringRules. *Journal of Statistical Software*, 90(12):1–37, 2019.

Wentao Li, Baoxiang Pan, Jiangjiang Xia, and Qingyun Duan. Convolutional neural network-based statistical post-processing of ensemble precipitation forecasts. *Journal of Hydrology*, 605:127301, 2022.

Jakob W. Messner, Georg J. Mayr, and Achim Zeileis. Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145(1):137–147, 2017.

Stephan Rasp and Sebastian Lerch. Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review*, 146(11):3885–3900, 2018.

Mark J. Rodwell, David S. Richardson, David B. Parsons, and Heini Wernli. Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, 99(5):1015–1026, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing.

Michael Scheuerer, Matthew B. Switanek, Rochelle P. Worsnop, and Thomas M. Hamill. Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Monthly Weather Review*, 148(8):3489–3506, 2020.

Benedikt Schulz and Sebastian Lerch. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review*, 150(1):235–257, 2022.

Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

Stéphane Vannitsem, Daniel S. Wilks, and Jakob Messner. *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 2018.

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R. Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, Zied Ben Bouallègue, Jonas Bhend, Markus Dabernig, Lesley De Cruz, Leila Hieta, Olivier Mestre, Lionel Moret, Iris Odak Plenković, Maurice Schmeits, Maxime Taillardat, Joris Van den Bergh, Bert Van Schaeybroeck, Kirien Whan, and Jussi Ylhaisi. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):E681 – E699, 2021.

Simon Veldkamp, Kirien Whan, Sjoerd Dirksen, and Maurice Schmeits. Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Monthly Weather Review*, 149(4): 1141–1152, 2021.

Daniel S Wilks. *Statistical Methods in the Atmospheric Sciences*. Elsevier Academic Press, 3rd edition, 2011.

A Appendix: Supplemental material

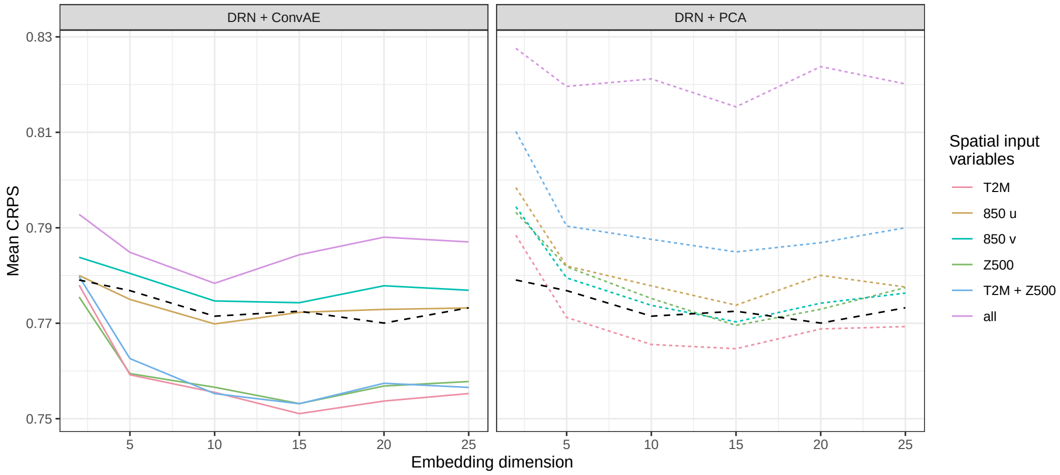
A.1 Model specification details and sensitivity experiments

ConvAE models

The final architecture of the ConvAE model shown in Figure 1 was chosen based on preliminary experiments on the training dataset. The encoder part consists of a sequence of convolutional (with 16, 8 and 4 filters) and max-pooling layers. The 2D-convolution layers use 3×3 kernels with a stride of 1, zero-padding and a ReLU activation. The number of filters and the kernel size were chosen to balance computational costs and representation quality, but the results were generally fairly robust to changes in these parameters. The max-pooling layers use a window size of 3×3 and a stride of 1. In the central dense encoding layer, we apply a linear activation function and ReLU activations in the neighboring dense layer. Following the standard practice in the extant literature (e.g., Ronneberger et al., 2015) the decoder part of the ConvAE model is based on 2D-transposed convolution layers, here with 4, 8 and 16 filters, a decoder kernel of size 9×9 and ReLU activations. The final output is obtained via a 2D-convolution layer with sigmoid activation. The model is trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. To prevent overfitting, we set the maximum number of epochs to 100 and apply early stopping with a patience of 10. Thereby, data from 2007–2014 is used for training, and data from 2015 as a validation dataset.

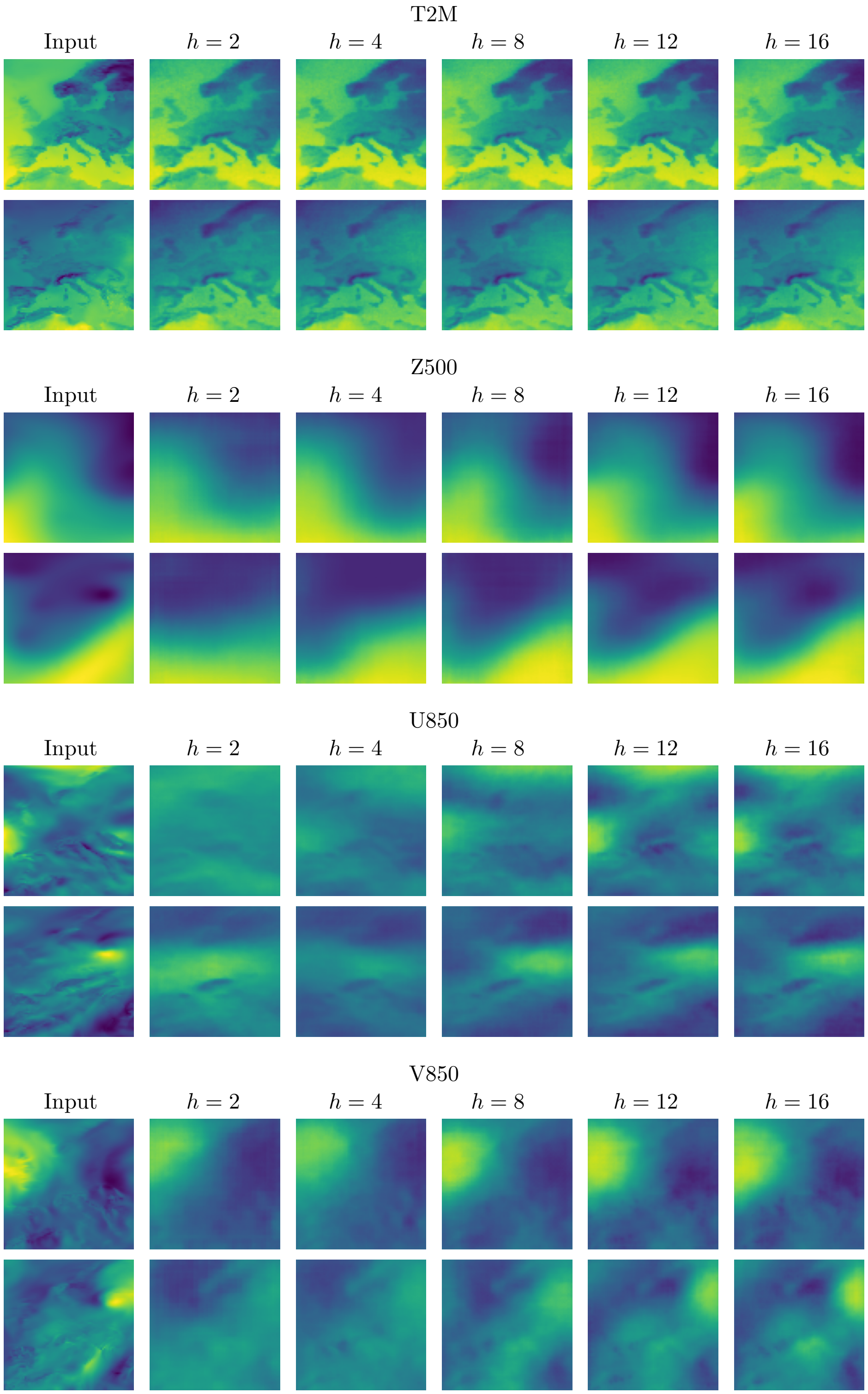
DRN and DRN+ConvAE/PCA models

The DRN and the DRN+ConvAE/PCA models share a common architecture to enable a fair comparison and a direct investigation of the effect of including spatial inputs. All models use two hidden layers with 100 nodes and ReLU activations. While including a second hidden layer deviates from the choices in Rasp & Lerch (2018), we found that this has only a negligible effect on the performance of the DRN model, but does improve the models with spatial inputs. The models are trained using the Adam optimizer with a learning rate of 0.002 for a maximum of 100 epochs, and early stopping with a patience of 10 is applied to prevent overfitting. We produce an ensemble of NN models by repeating the model estimation 10 times, and aggregate the predictions by averaging the distribution parameters. Data from 2007–2014 is used for training, and data from 2015 for validation purposes. An important tuning parameter is the dimension of the station embeddings. While Rasp & Lerch (2018) use two-dimensional embeddings, subsequent research demonstrated the usefulness of choosing larger values (e.g., Bremnes, 2020; Schulz & Lerch, 2022). We chose an embedding dimension of 15 for all models. Supplementary Figure 1 shows the mean CRPS as functions of the embedding dimension and indicates that for the models with added spatial inputs, choosing larger than 2 improves the forecast performance, wheres the effects on the DRN model are relatively minor.



Supplementary Figure 1: Mean CRPS over the test set as a function of the embedding dimension for the DRN+ConvAE and DRN+PCA models, with different types of spatial inputs and $h = 2$. The black dashed line indicates the mean CRPS of the DRN model without spatial inputs.

A.2 EXEMPLARY CONVAE RECONSTRUCTIONS

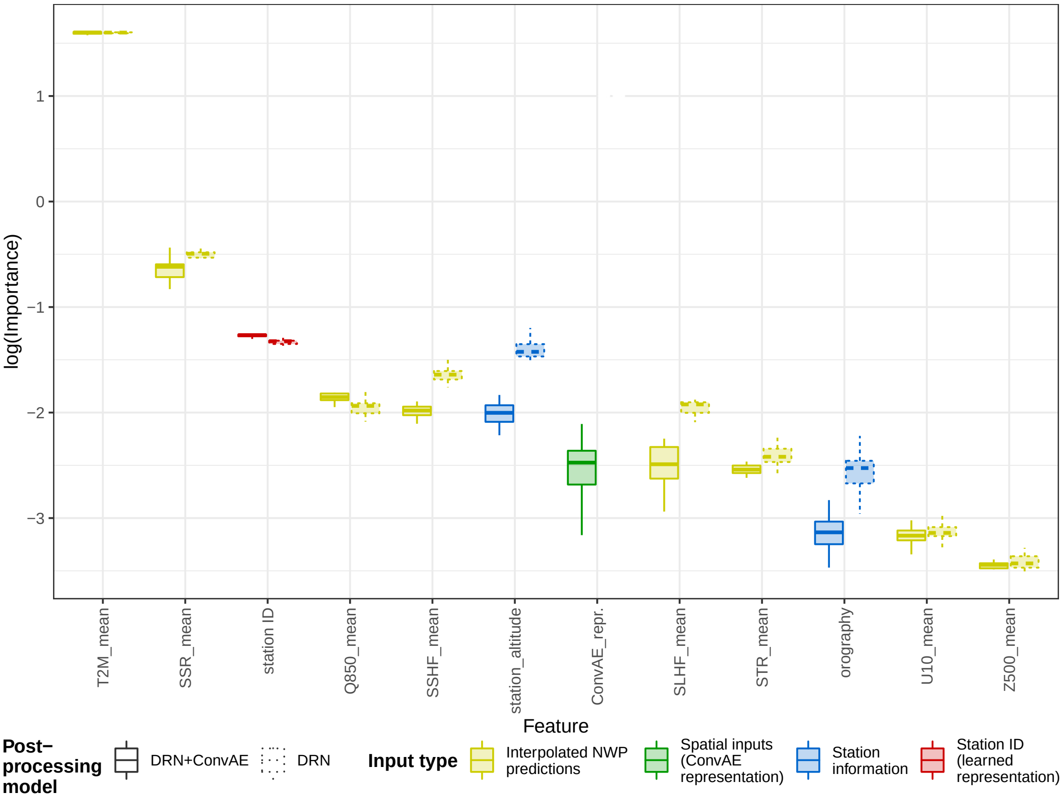


Supplementary Figure 2: Exemplary ConvAE reconstructions of randomly selected examples from the test dataset for different values of the encoding dimension h.

A.3 Feature importance

Supplementary Figure 3 shows the permutation-based feature importance of the 12 most important predictors of the DRN+ConvAE and the DRN model. To compute feature importances, we follow Rasp & Lerch (2018) and Schulz & Lerch (2022), and measure the decrease in terms of the CRPS in the test set when randomly permuting a single input feature, using the mean CRPS of the respective model based on unpermuted input features as reference.

Overall, the rankings among the most important features are relatively consistent among the two models, but a decreased importance can be observed for the DRN+ConvAE model for most of the features compared to the DRN model, most notably for the station altitude and orography. The feature importance of the ConvAE representations of the spatial T2M inputs ranks seventh on average, but shows a notably larger variability over repetitions of the model fitting procedure compared to the other inputs.



Supplementary Figure 3: Permutation-based feature importances of the 12 most importance predictors of the DRN+ConvAE (with T2M inputs and $h = 2$) and DRN model shown on a logarithmic scale. The input features are colored by type according to the illustration in Figure 1 in the main text, see Rasp & Lerch (2018) for abbreviations of the interpolated NWP variables. The boxplots indicate the variability of the importances across the 10 repetitions of the model fitting procedure.