Deep learning models for distributional regression

Sebastian Lerch (Karlsruhe Institute of Technology)

WU Wien, February 2020

Outline

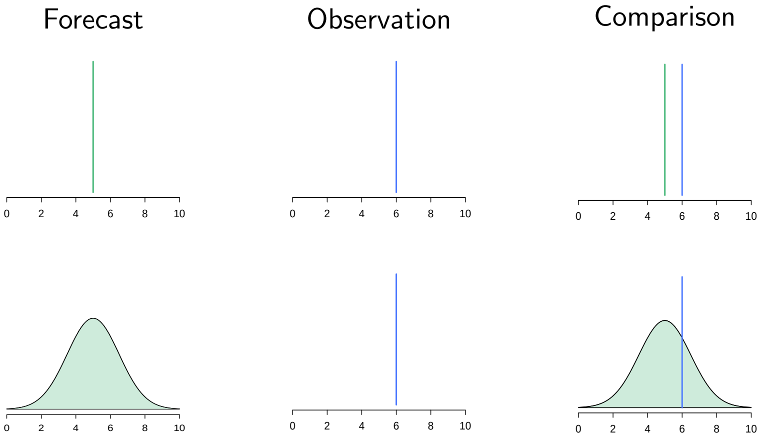
- Probabilistic forecasting and comparative model assessment
- Motivation: Post-processing ensemble weather simulations
- Neural networks for distributional regression
- Advanced machine learning methods for incorporating complex sources of information

Probabilistic forecasting

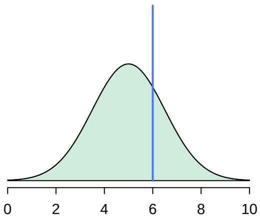
Model predictions should be probabilistic (given as a parametric or simulation-based probability distribution) to

- quantify inherent uncertainty
- allow for optimal decision making by obtaining target functionals (mean, quantiles, ...) of the predictive distributions
- meet increasing popularity and requests across disciplines, in particular in economics and environmental sciences.

Deterministic and probabilistic predictions



Evaluation of probabilistic forecasts: Proper scoring rules



A (negatively oriented) proper scoring rule is any function

$$S(F, y)$$

such that for all F, G,

$$\mathbb{E}_{Y \sim G} S(G, Y) \leq \mathbb{E}_{Y \sim G} S(F, Y).$$

Popular examples include

the logarithmic score

$$\mathrm{LogS}(F, y) = -\log(f(y))$$

the continuous ranked probability score

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 dz$$

Proper scoring rules as tools for model estimation

Proper scoring rules provide useful tools for parameter estimation in an M-estimation framework: Determine

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} S(F_\theta, y_i).$$

LogS yields maximum likelihood (ML) estimation, the CRPS provides a robust alternative.

Computational tools: Efficient implementations for parametric and simulation-based predictive models for optimization and large scale evaluation: `R` package `scoringRules`.

Jordan, A., Krüger, F. and Lerch, S. (2019)
Evaluating probabilistic forecasts with scoringRules.
Journal of Statistical Software, 90, 1–37.

Outline

- Probabilistic forecasting and comparative model assessment
- Motivation: Post-processing ensemble weather simulations
- Neural networks for distributional regression
- Advanced machine learning methods for incorporating complex sources of information

Numerical weather prediction

Modern weather forecasts rely on physical numerical weather prediction (NWP) models of atmospheric processes.

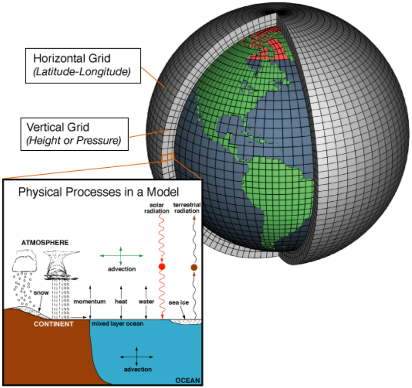


Image source: NOAA[1]

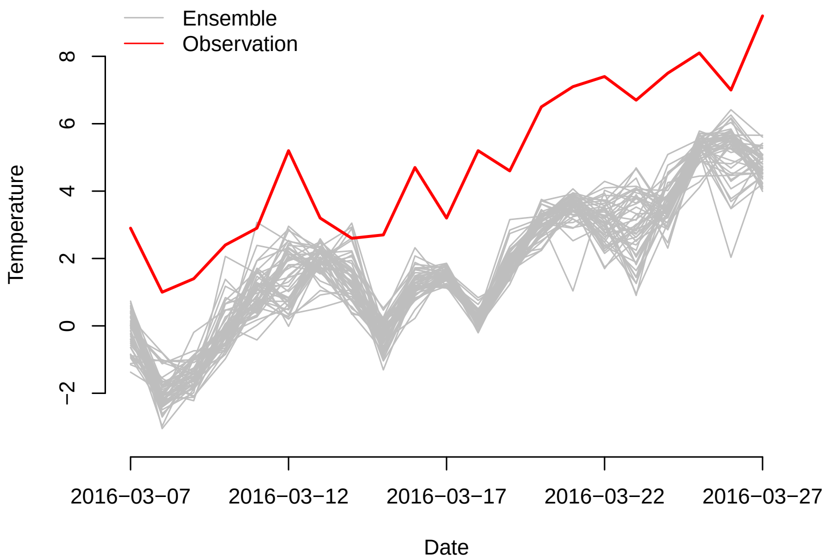
However, there are major sources of uncertainty (initial conditions, physical models).

Ensemble simulations seek to quantify uncertainty and provide probabilistic forecasts.

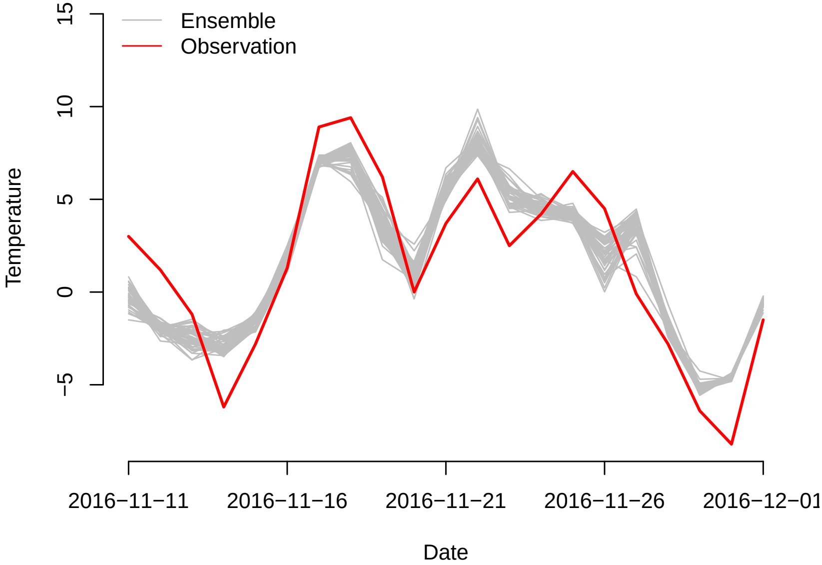
Despite continued improvements, ensemble forecasts are subject to model biases and lack calibration.

[1] https://celebrating200years.noaa.gov/breakthroughs/climate_model/AtmosphericModelSchematic.png

Example: Ensemble forecasts of temperature



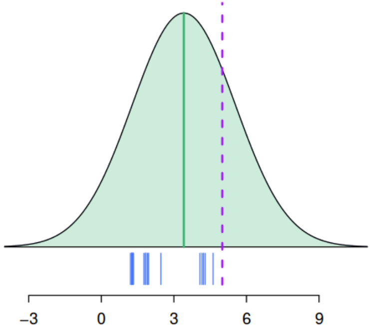
Example: Ensemble forecasts of temperature



Statistical post-processing of ensemble forecasts

Ensemble simulations typically fail to accurately quantify model uncertainty and require calibration via statistical post-processing.

Example: Temperature: Using ensemble predictions of temperature as input, the post-processed forecast takes the form of a Gaussian distribution.



$$y|\boldsymbol{X}^{\text{t2m}} \sim \mathcal{N}_{(\mu,\sigma)},$$

$$\mu = a + b \cdot \text{mean}(\boldsymbol{X}^{\text{t2m}})$$
$$\sigma = c + d \cdot \text{sd}(\boldsymbol{X}^{\text{t2m}})$$

Distributional regression models for post-processing

Model probability distribution of target variable y given ensemble model output $\boldsymbol{X}$ by a parametric distribution $F_{\boldsymbol{\theta}}$,

$$y|\boldsymbol{X} \sim F_{\boldsymbol{\theta}}, \qquad \text{where} \qquad \boldsymbol{\theta} = g(\boldsymbol{X}).$$

Limitations of fully parametric approaches:

- requires choice of link function g connecting predictors $\boldsymbol{X}$ and distribution parameters $\boldsymbol{\theta}$
  - difficult to specify functional form of dependencies if many possible predictors are available
- requires estimation of parameters of g
  - global (using all training data) or local (location-specific) models?
- requires choice of parametric model F_θ

Advanced benchmark models

Including additional predictors is not straightforward. To avoid overfitting, predictor selection strategies are required.

- Gradient boosting approach (EMOS-loc-bst model) proposed by Messner et al. (2017, MWR): Assume

$$(\mu, \sigma) = \left(\boldsymbol{X}^\mathsf{T} \boldsymbol{\beta},\ \exp(\boldsymbol{X}^\mathsf{T} \gamma) \right),$$

  and iteratively update coefficient vector entries improving the current model fit most.

- Quantile regression forest (QRF) model proposed by Taillardat et al. (2016, MWR): Nonparametric quantile regression based on random forests. Quantile estimates are obtained from an ensemble of decision trees.

Have to be implemented as local models to achieve good forecasts.

Outline

- Probabilistic forecasting and comparative model assessment
- Motivation: Post-processing ensemble weather simulations
- Neural networks for distributional regression
- Advanced machine learning methods for incorporating complex sources of information

Neural networks for post-processing ensemble forecasts

Novel semi-parametric approach: Estimate distribution parameters θ directly by training a neural network to

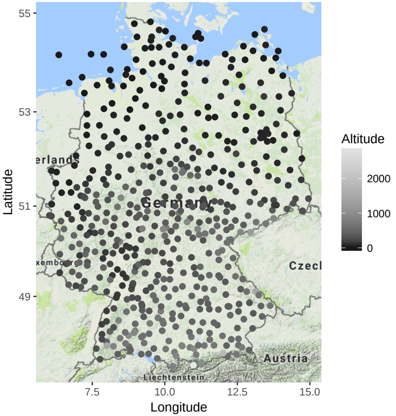
- learn arbitrary nonlinear relations in an automated, data-driven manner,
- generate local adaptivity in globally estimated models,
- gain meteorological insight from trained models.

Rasp, S. and Lerch, S. (2018)
Neural networks for post-processing ensemble weather forecasts,
Monthly Weather Review, 146, 3885–3900.

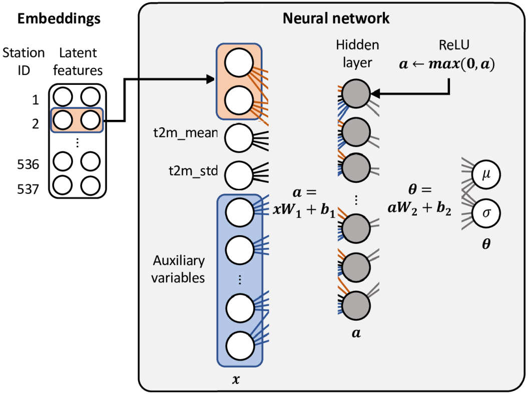
`Python`/`R` code available at `https://github.com/slerch/ppnn`.

Data

- data from 2007–2016
- 48 hours-ahead ECMWF 50-member ensemble forecasts of temperature (and 17 other variables)
- DWD station observations at 537 locations
- data from 2016 used as evaluation set
- two training datasets: 2015 and 2007–2015



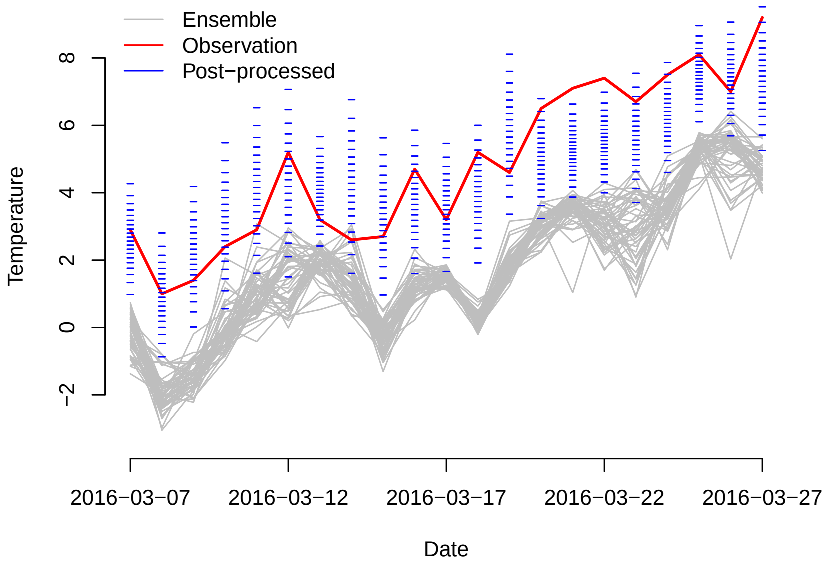
Neural networks for distributional regression



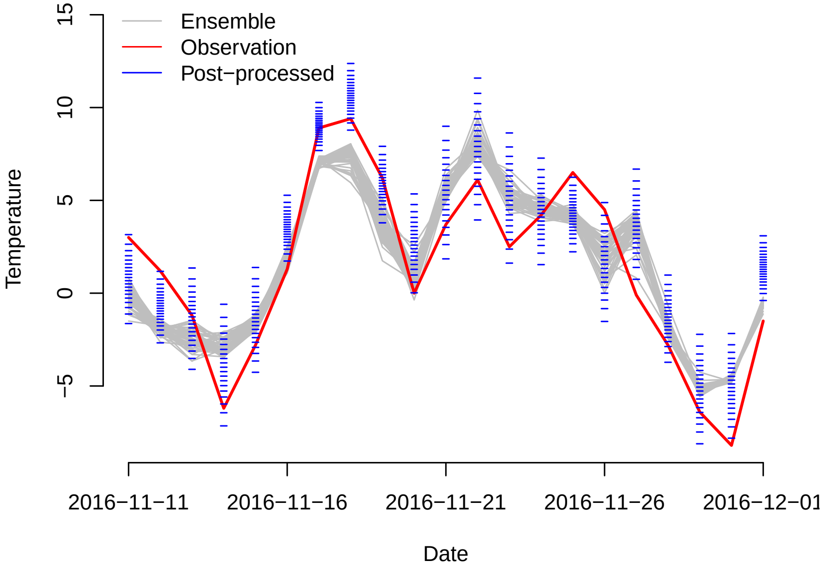
- Input: Predictor variables (NWP quantities, station characteristics).
- Output: Distribution parameters $\boldsymbol{\theta}$
- Embeddings generate local adaptivity.

Training via CRPS minimization (mathematically principled non-standard choice).

Example: Ensemble forecasts of temperature



Example: Ensemble forecasts of temperature

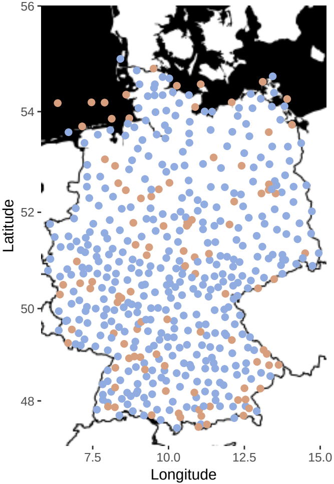


Overview of results

CRPS: Continuous ranked probability score, lower is better

Model	Description	Mean CRPS for training period 2015	Mean CRPS for training period 2007–2015
Raw ensemble		1.16	1.16
Benchmark post-processing methods			
EMOS-gl	Global EMOS	1.01	1.00
EMOS-loc	Local EMOS	0.90	0.90
EMOS-loc-bst	Local EMOS with boosting	0.85	0.80
QRF	Local quantile regression forest	0.95	0.81
Neural network models			
NN-aux-emb	Neural network with auxiliary predictors and station embeddings	**0.82**	**0.78**

Station-specific comparison of NN and benchmark models

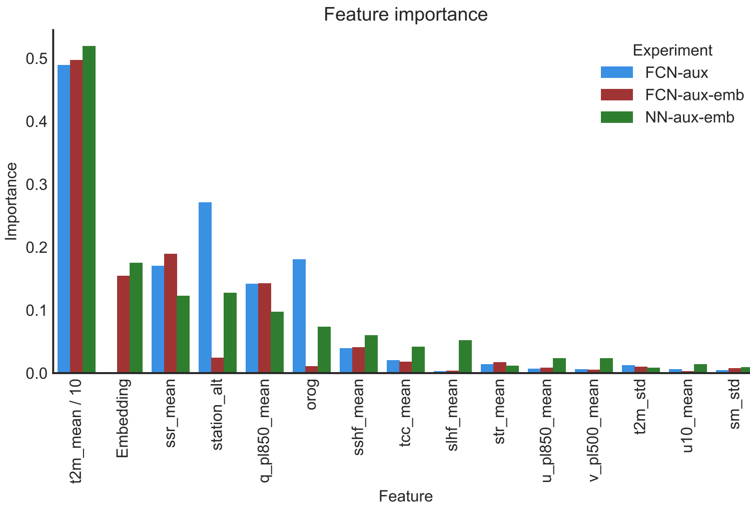


Station-specific best model is a NN model / benchmark model

NN models perform best at more than 80% of the stations.

Differences are statistically significant at a large fraction of stations.

Peeking into the black box of neural network models



Change in mean CRPS after permuting a single input variable according to a random permutation across stations and dates.

Outline

- Probabilistic forecasting and comparative model assessment
- Motivation: Post-processing ensemble weather simulations
- Neural networks for distributional regression
- Advanced machine learning methods for incorporating complex sources of information

Prospects of modern ML for distributional regression

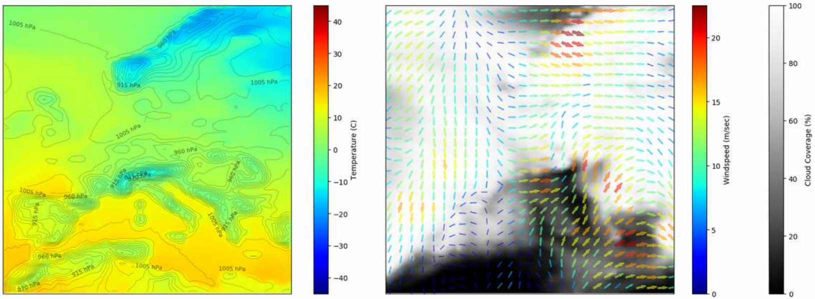
Modern AI methods provide unprecedented tools for data analysis and prediction.

In particular, machine learning can be useful for

- incorporating spatial, temporal and inter-variable information into model building and estimation,
- incorporating prior knowledge about underlying (e.g. physical) processes into models,
- flexible modelling of complex response distributions.

Spatial information

Ensemble forecasts are gridded 2D fields of forecasts of weather variables. Thus far, those were interpolated to station locations.

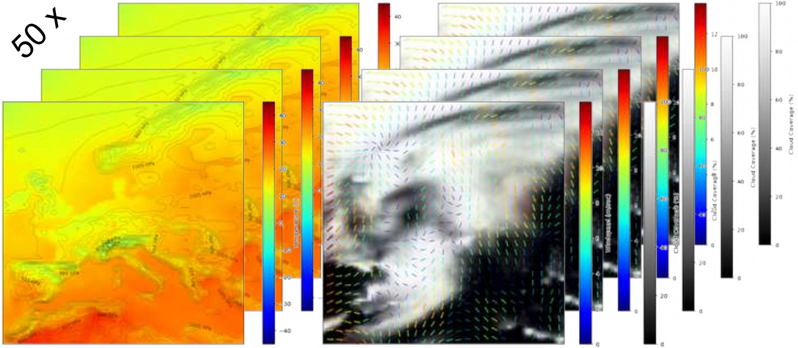


Gridded ECMWF forecasts over Europe (0.5° resolution, 81 × 81 pixels)

However, large-scale spatial structure and predictability information (e.g., 'weather regimes') get lost in the interpolation step.

Ensemble information

Ensemble members provide 50 physically coherent forecasts of weather variables. Thus far, only mean and standard deviation of (interpolated) ensemble forecasts were used.

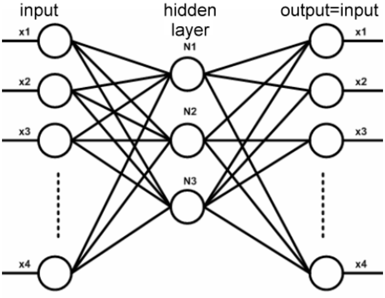


Possibly important uncertainty information might get lost by the use of summary statistics.

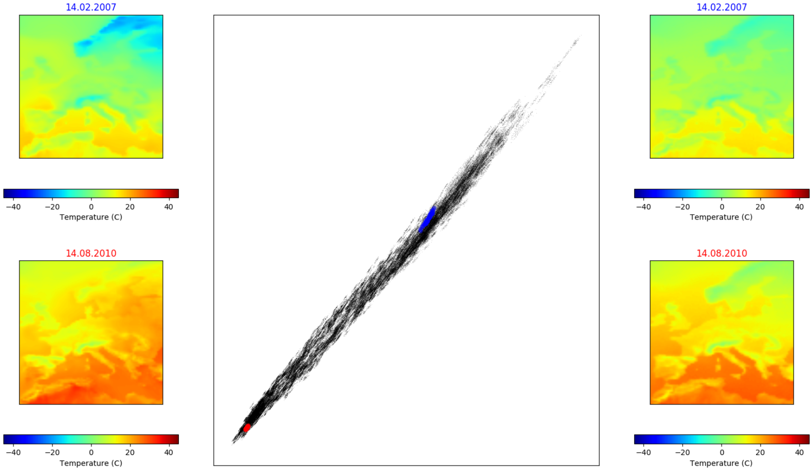
Deep autoencoders for dimensionality reduction

Specific NN architectures to find compact representation of inputs (unsupervised) by

- training the network to re-create its own inputs
- creating a bottleneck by using fewer hidden units than inputs

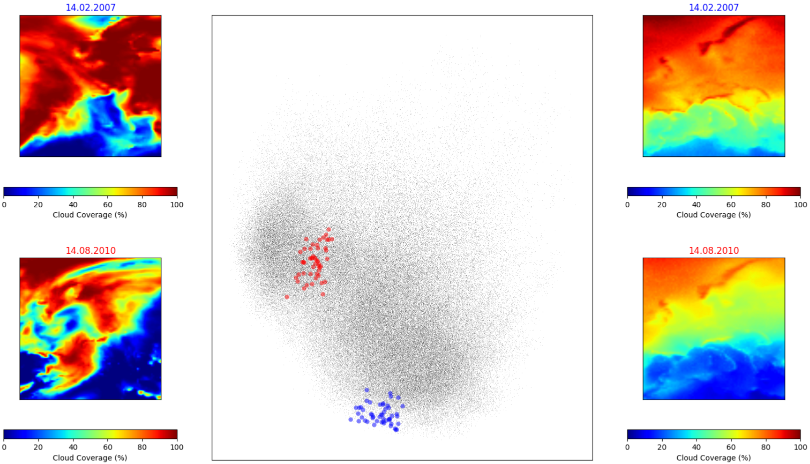


Projections of ensemble forecasts (temperature)



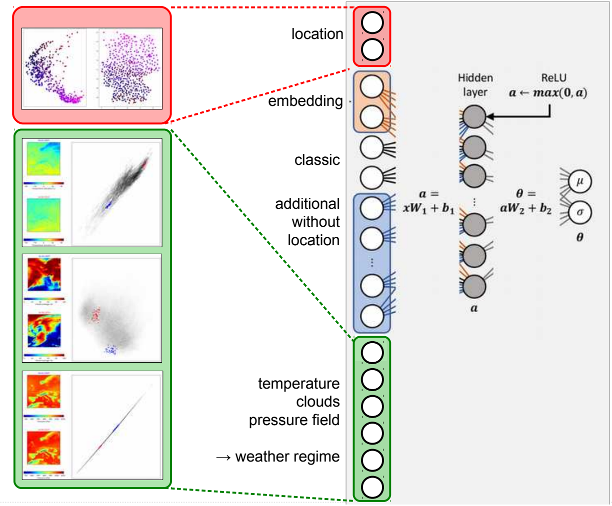
Left: Example input forecast fields from two days.
Middle: Ensemble members in projected space (blue: top, red: bottom).
Right: Reconstructed fields.

Projections of ensemble forecasts (cloud cover)



Left: Example input forecast fields from two days.
Middle: Ensemble members in projected space (blue: top, red: bottom).
Right: Reconstructed fields.

Autoencoder representations as additional NN-input



Preliminary results suggest improvements in mean CRPS (0.78 → 0.76).

Ongoing joint work with Kai Polsterer and Antonio D'Isanto.

Incorporating physical knowledge into ML models

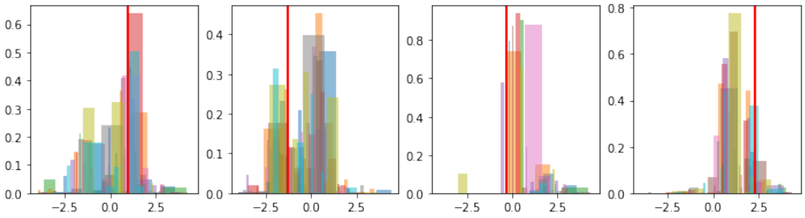
Modern AI methods provide new approaches to better understand and utilize interactions of domain knowledge and statistics, which is key to improving forecasting systems and optimize predictions.

In the context of ensemble post-processing, examples include

- incorporating predictability information contained in large-scale weather patterns ('weather regimes'), e.g. utilizing indicators as predictors
- stratified model estimation by objectively identified and meteorologically meaningful dynamic subregions of storms (PhD project of Benedikt Schulz)

Neural networks for nonparametric distributional regression

The choice of a suitable parametric forecast distribution F_θ remains a challenge for parametric approaches.



NN-based nonparametric distributional regression methods may allow to flexible model complex response distributions.

Ongoing joint work with Stephan Rasp, M.Sc. thesis by Marvin Bischoff on electric load forecasting.

Summary

- semi-parametric distributional regression models based on neural networks
- flexible, automated and data-driven modelling of nonlinear relations between predictors and distribution parameters
- perform better than state of the art approaches and allow to gain meteorological insight from trained models
- compressing complex spatial data might improve performance and add to interpretability

Rasp, S. and Lerch, S. (2018)
Neural networks for post-processing ensemble weather forecasts,
Monthly Weather Review, 146, 3885–3900.

Python/R code available at `https://github.com/slerch/ppnn`.