Probabilistic weather prediction: From ensembles to neural networks

Sebastian Lerch

Karlsruhe.ai TechTalk, January 2018

Heidelberg Institute for Theoretical Studies
HITS

Karlsruhe Institute of Technology

WAVES TO WEATHER

Questions you can – hopefully – answer after my talk

- Why should predictions be probabilistic?
- How can probabilistic forecasts be evaluated, and what can mathematical statistics contribute?
- How are modern weather forecasts produced, and what is the role of statistics?
- Isn't this supposed to be an AI meetup?

Overview

- Making and evaluating probabilistic forecasts
- Weather forecast ensembles and statistical post-processing
- Multivariate copula-based post-processing approaches
- Neural network approaches

Overview

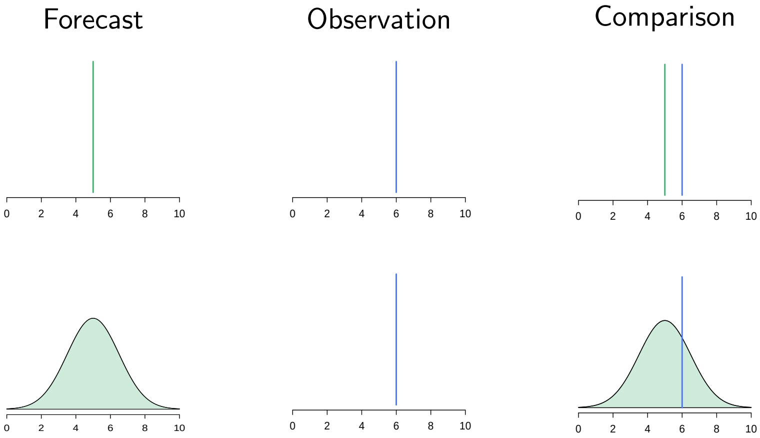
- Making and evaluating probabilistic forecasts
- Weather forecast ensembles and statistical post-processing
- Multivariate copula-based post-processing approaches
- Neural network approaches

Probabilistic forecasts

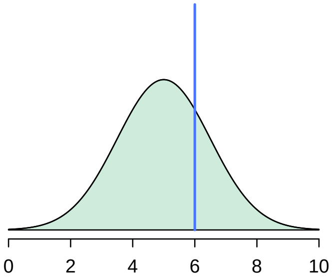
Probabilistic forecasts, i.e., forecasts in the form of probability distributions over future quantities or events,

- provide information about inherent uncertainty
- allow for optimal decision making by obtaining deterministic forecasts as target functionals (mean, quantiles, ...) of the predictive distributions
- have become increasingly popular across disciplines: meteorology, hydrology, seismology, economics, finance, demography, political science, ...

Probabilistic vs. point forecasts



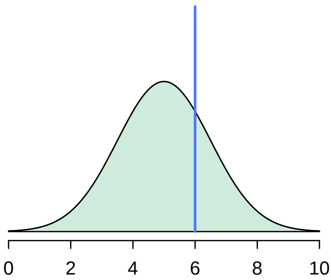
What is a good probabilistic forecast?



The goal of probabilistic forecasting is to maximize the sharpness of the predictive distribution subject to calibration.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) **Probabilistic forecasts, calibration and sharpness**. *Journal of the Royal Statistical Society Series B*, 69, 243–268.

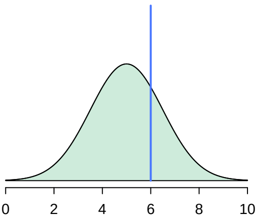
Calibration and sharpness



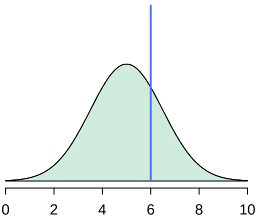
Calibration: Compatibility between the forecast and the observation; joint property of the forecasts and observations

Sharpness: Concentration of the forecasts; property of the forecasts only

Evaluation of probabilistic forecasts: Proper scoring rules



Evaluation of probabilistic forecasts: Proper scoring rules



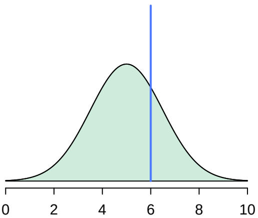
A (negatively oriented) proper scoring rule is any function

$$S(F, y)$$

such that for all F, G,

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

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,

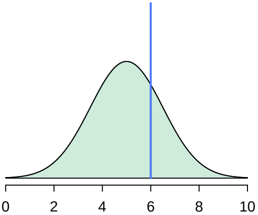
$$\mathbb{E}_{Y \sim G} S(G, Y) \leq \mathbb{E}_{Y \sim G} S(F, Y).$$

Popular examples include

the logarithmic score

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

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

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

Advertisement: R package scoringRules

Joint work with Alexander Jordan (HITS) and Fabian Krüger (Heidelberg University).

Aim: Provide a user-friendly toolbox

- computation of popular scoring rules for simulated samples,
- and for a variety of parametric distributions, including many previously unavailable closed-form expressions of the CRPS

Available from CRAN and
https://github.com/FK83/scoringRules.

For more information, see

Jordan, A., Krüger, F. and Lerch, S. (2017) **Evaluating probabilistic forecasts with the R package scoringRules**. https://arxiv.org/abs/1709.04743.

Overview

- Making and evaluating probabilistic forecasts
- Weather forecast ensembles and statistical post-processing
- Multivariate copula-based post-processing approaches
- Neural network approaches

Numerical weather prediction (NWP)

Atmospheric processes are subject to physical laws that can be described by differential equations.

NWP: Simulation of these processes to calculate how the weather will evolve starting from its present state.

[ ]

http://www.ecmwf.int/sites/default/files/parametrization_0.png

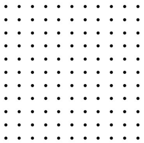
NWP models calculate future states of the atmosphere

- for several variables of the atmosphere
- on a spatial grid around the globe, discretized in time

state of the atmosphere

at time t

initial conditions



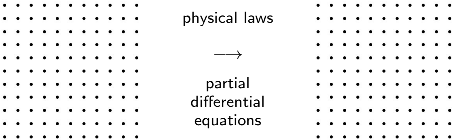
NWP models calculate future states of the atmosphere

- for several variables of the atmosphere
- on a spatial grid around the globe, discretized in time

state of the atmosphere

at time t | at time $t+1$

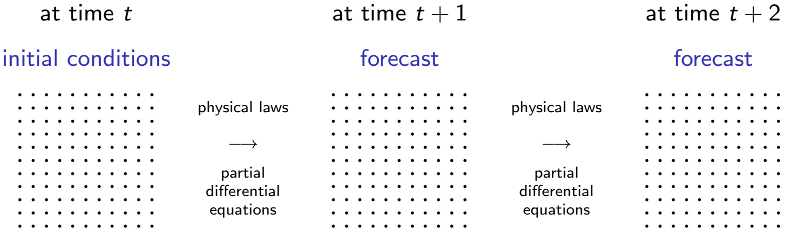
initial conditions | forecast



NWP models calculate future states of the atmosphere

- for several variables of the atmosphere
- on a spatial grid around the globe, discretized in time

state of the atmosphere



Success of NWP

approximate gain of one day in predictability per decade

[ ]

https://www.nature.com/articles/nature14956/figures/1

Bauer, P., Thorpe, A. and Brunet, G. (2015) **The quiet revolution of numerical weather prediction**. *Nature*, 525, 47–55.

Probabilistic weather forecasts

However, there are major sources of uncertainty, including uncertainty about initial conditions and physical models, e.g.,

- observations for many variables are sparse or not accurately measured (*butterfly effect*!)
- many processes happen on on sub-grid scale

Probabilistic weather forecasts

However, there are major sources of uncertainty, including uncertainty about initial conditions and physical models, e.g.,

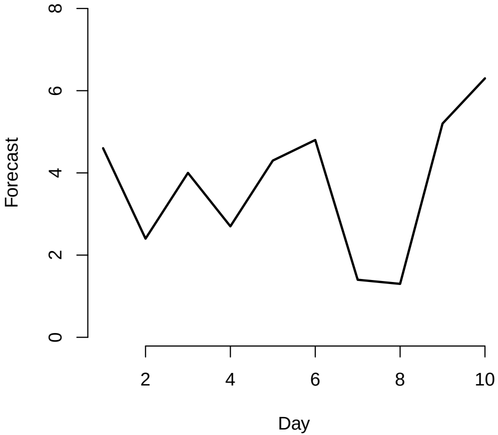
- observations for many variables are sparse or not accurately measured (*butterfly effect*!)
- many processes happen on on sub-grid scale

Over the last three decades: Radical culture change towards carefully designed ensembles of NWP model runs to quantify uncertainty

- each ensemble member is a single-valued, deterministic forecast using an NWP model
- forecasts differ with respect to initial conditions and/or model formulation

Example: Wind speed forecasts at Frankfurt airport

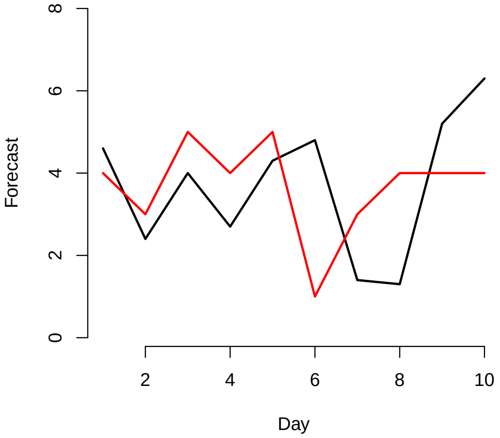
European Centre for Medium-Range Weather Forecasting (ECMWF) predictions, initialized July 1, 2012



single unperturbed model run

Example: Wind speed forecasts at Frankfurt airport

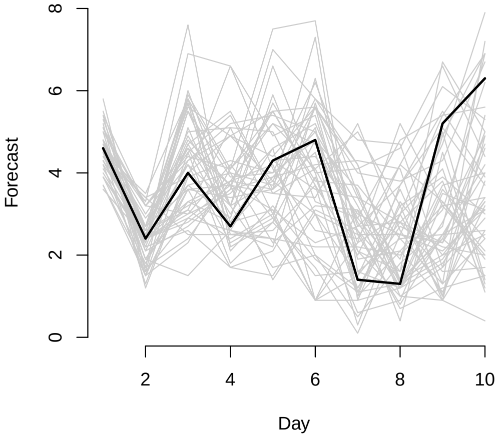
European Centre for Medium-Range Weather Forecasting (ECMWF) predictions, initialized July 1, 2012



single unperturbed model run, observation

Example: Wind speed forecasts at Frankfurt airport

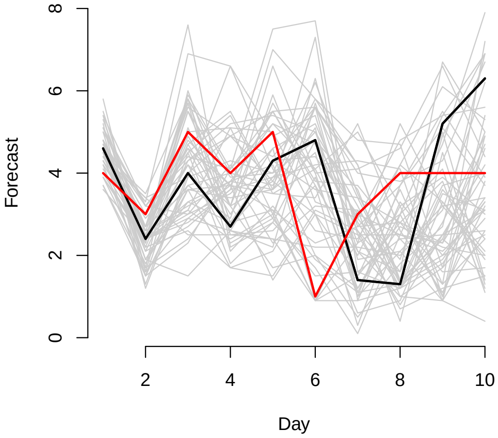
European Centre for Medium-Range Weather Forecasting (ECMWF) predictions, initialized July 1, 2012



single unperturbed model run, ensemble

Example: Wind speed forecasts at Frankfurt airport

European Centre for Medium-Range Weather Forecasting (ECMWF) predictions, initialized July 1, 2012



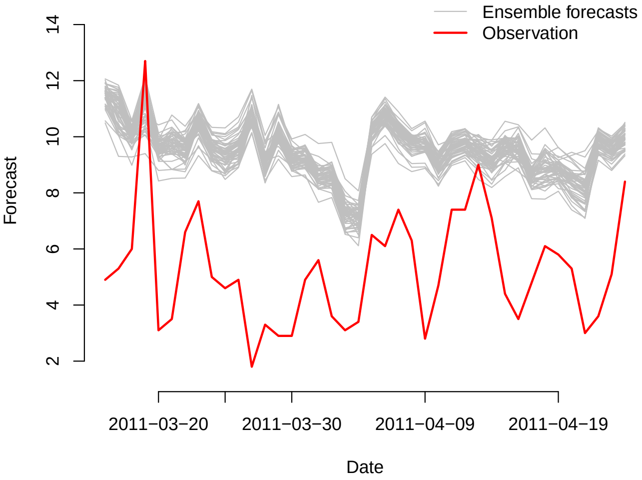
single unperturbed model run, ensemble, observation

Deficiencies of ensemble forecasts

Despite their undisputed success, ensemble forecasts typically fail to represent the full model uncertainty.

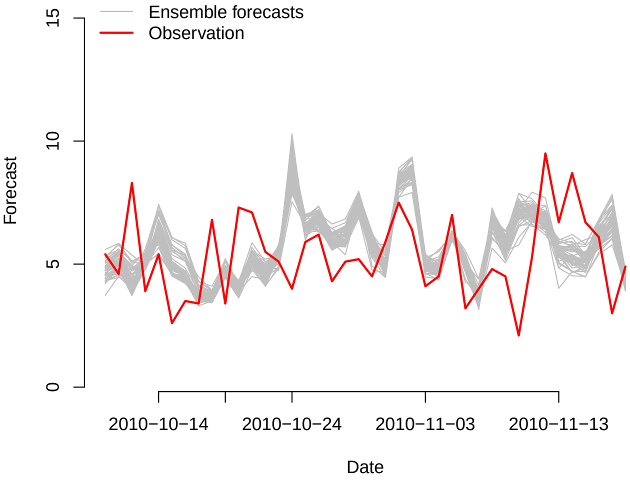
In particular, they are usually subject to model biases and lack calibration.

Example: Wind speed forecasts at Frankfurt airport



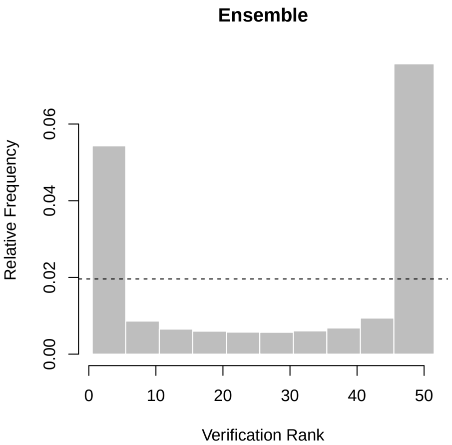
1-day ahead ECMWF ensemble forecasts of daily maximum wind speed

Example: Wind speed forecasts at Frankfurt airport



1-day ahead ECMWF ensemble forecasts of daily maximum wind speed

Under-dispersion of ECMWF wind speed forecasts



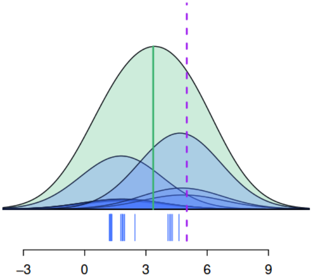
If the ensemble were a random sample from the true distribution of the weather quantity, the rank distribution of the observation when pooled with the ensemble would be uniform.

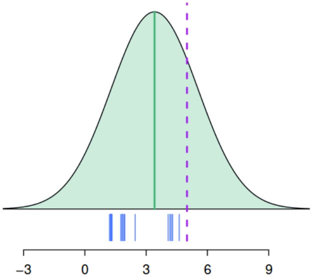
Statistical post-processing of ensemble forecasts

Ensemble forecasts thus require some form of statistical post-processing.

- basic idea: exploit structure in past forecast-observation pairs to correct for systematic errors in the model output
- approach depends on the availability of suitable training sets, consisting of past forecast-observation pairs to estimate statistical models
- post-processing combines numerical weather models and statistical regression modeling

Statistical post-processing approaches: Overview





1. Bayesian model averaging (BMA): each ensemble member is associated with a kernel function, with weight that reflects that member's skill
2. Ensemble model output statistics (EMOS) or non-homogeneous regression (NR): fits a single, parametric predictive distribution using summary statistics from the ensemble.

Bayesian model averaging (BMA)

Let y denote the weather variable of interest, and $x_1, \ldots, x_m$ the corresponding ensemble member forecasts.

The BMA predictive distribution is a weighted mixture

$$y|x_1, \ldots, x_m \sim \sum_{i=1}^{m} w_i \, f(y|x_i)$$

where w_i are weights that sum to 1, and $f(y|x_i)$ is a suitable parametric density that depends on the ensemble member x_i

e.g. normal distributions for temperature and pressure, or gamma distributions for wind speed.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) **Using Bayesian model averaging to calibrate forecast ensembles**. *Monthly Weather Review*, 133, 1155–1174.

Non-homogeneous regression (NR)

Fit a single, parametric predictive distribution,

$$y|x_1, \ldots, x_m \sim f(y|x_1, \ldots, x_m),$$

the parameters of which depend on the ensemble forecasts through suitable link functions.

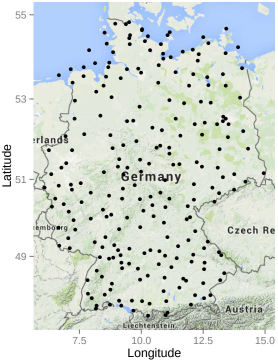
For example, in case of temperature or pressure f can be chosen to be a normal distribution.

For other weather variables such as wind speed or precipitation, the choice of a parametric family is less obvious.

Gneiting, T., Raftery, A. E., Westveld III, A. H. and Goldman, T. (2005) **Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation**. *Monthly Weather Review*, 133, 1098–1118.

Exemplary case study

- 1-day ahead forecasts and observations of daily maximum wind speed
- ECMWF ensemble forecasts, 50 members
- 228 observation stations over Germany
- evaluation period:
  May 1, 2010 – April 30, 2011
- $> 80\,000$ individual forecast cases



Lerch, S. and Thorarinsdottir, T. L. (2013) **Comparison of non-homogeneous regression models for probabilistic wind speed forecasting**. *Tellus A*, 65, 21206.

Non-homogeneous regression models for wind speed

Standard NR model for wind speed: truncated normal (TN) distribution

$$y|x_1, \ldots, x_m \sim \mathcal{N}_{[0,\infty)}(\mu, \sigma^2),$$

where

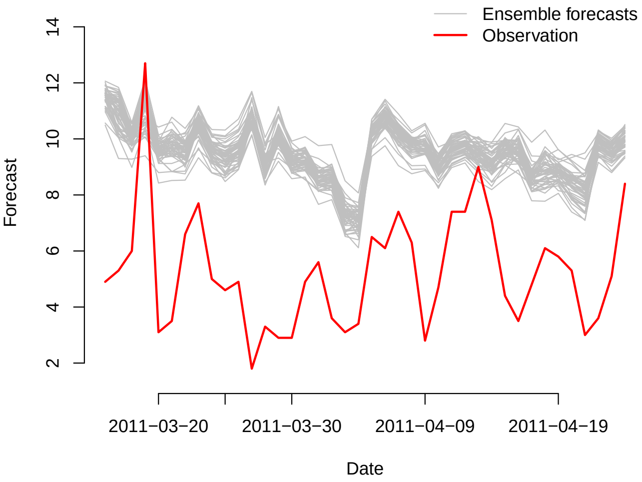
$$\mu = a + b\,\bar{X} \quad \text{and} \quad \sigma^2 = c + d\,S^2,$$

here, $\bar{X}$ denotes the average ensemble forecast, and S^2 is the ensemble variance.

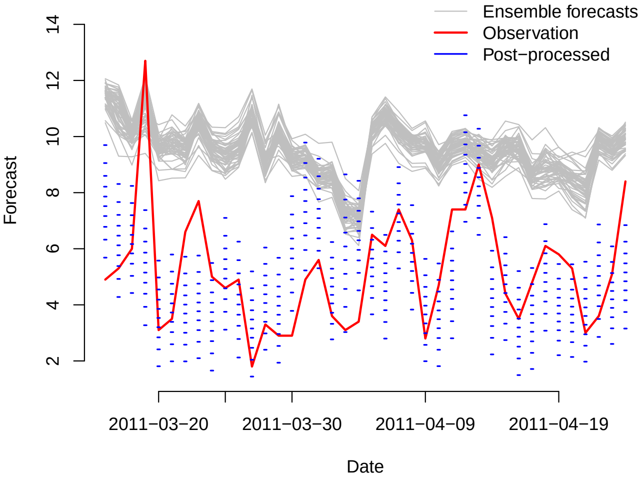
Parameters a, b, c, d are estimated over a rolling training period consisting of past pairs of forecasts and observations by numerically minimizing the mean CRPS.

Thorarinsdottir, T. L. and Gneiting, T. (2010) **Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression**. *Journal of the Royal Statistical Society Series A*, 173, 371–388.

Results: Wind speed forecasts at Frankfurt airport

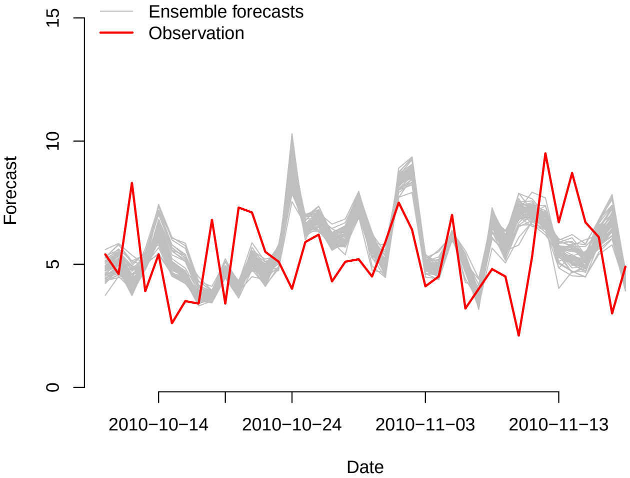


Results: Wind speed forecasts at Frankfurt airport

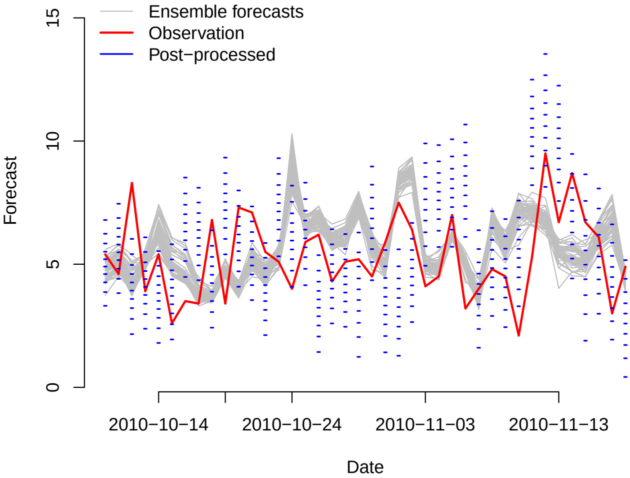


Post-processing removes the bias...

Results: Wind speed forecasts at Frankfurt airport

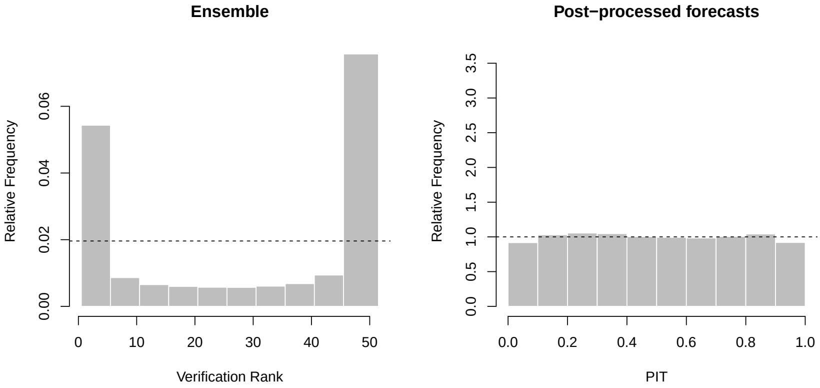


Results: Wind speed forecasts at Frankfurt airport



... and corrects for the under-dispersion.

Calibration



Under-dispersion of the ensemble forecasts is corrected by post-processing.

Research topics in post-processing

- choice of a suitable parametric family (e.g., skewed, heavy-tailed distributions are more appropriate for modeling wind speed)
- link functions to connect distribution parameters and summary statistics of ensemble forecasts
- utilization of covariates or additional meteorological information
- choice of suitable training sets (local or regional / bias-variance dilemma)
- handling of extreme events
- model development for specific applications and integration into operational use (e.g. in air traffic management)
- modeling of multivariate dependencies

Overview

- Ensembles and statistical post-processing
- Case study: Wind speed
- Multivariate copula-based post-processing approaches
- Challenges and future directions

Accounting for dependencies

NR and BMA apply to a single weather variable at a single location and a single look-ahead time.

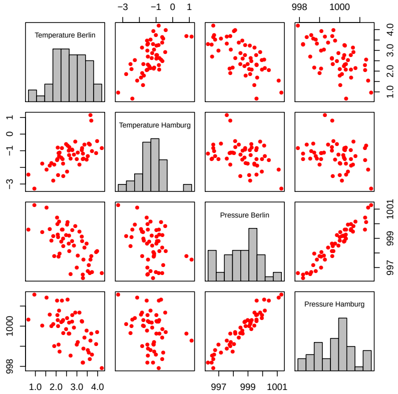
Individually post-processed distributions thus fail to account for multivariate dependence structures.

However, there is a need to develop post-processing techniques that yield physically realistic probabilistic forecasts of spatio-temporal weather trajectories.

Key applications include air traffic management, ship routeing, and renewable energy predictions.

Example

24-hour ahead ECMWF ensemble forecast of surface temperature and pressure at Berlin and Hamburg valid November 8, 2011 before and after NR post-processing



Example

24-hour ahead ECMWF ensemble forecast of surface temperature and pressure at Berlin and Hamburg valid November 8, 2011 before and after NR post-processing

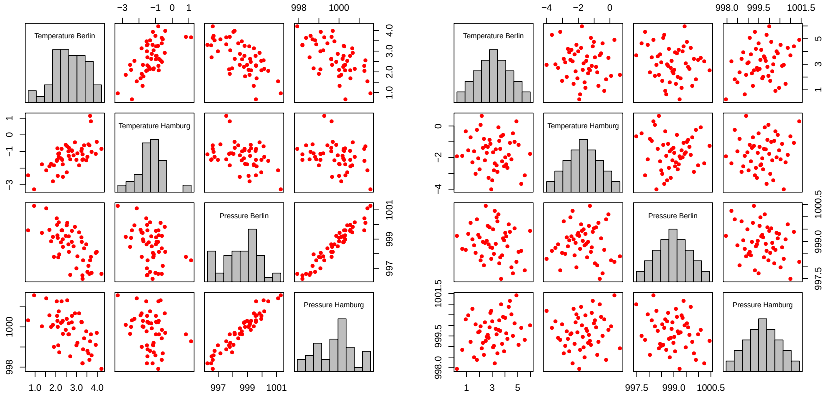


figure courtesy of Roman Schefzik

Sklar's theorem and copula-based approaches

Individual post-processing results in univariate probabilistic forecasts F_l for each $l = 1, \ldots, L$.

We seek a physically realistic and consistent multivariate predictive CDF, F, with margins $F_l, l = 1, \ldots, L$.

Sklar's theorem allows to connect F to the margins F_l via a copula C (a multivariate CDF with standard uniform margins),

$$F(y_1, \ldots, y_L) = C(F_1(y_1), \ldots, F_L(y_L)).$$

Sklar's theorem and copula-based approaches

Individual post-processing results in univariate probabilistic forecasts F_l for each $l = 1, \dots, L$.

We seek a physically realistic and consistent multivariate predictive CDF, F, with margins $F_l, l = 1, \dots, L$.

Sklar's theorem allows to connect F to the margins F_l via a copula C (a multivariate CDF with standard uniform margins),

$$F(y_1, \dots, y_L) = C(F_1(y_1), \dots, F_L(y_L)).$$

A suitable copula has to be chosen and fitted.

- for small L, or if specific structures can be exploited, parametric or semi-parametric copulas can be utilized.
- for large L and in general situations, non-parametric approaches based on empirical copulas can be used, ensemble copula coupling and the Schaake shuffle are attractive options.

Ensemble copula coupling

Given an ensemble of size m for weather variables $Y_l, l = 1, \ldots, L$, ECC proceeds in three steps:

1. Univariate post-processing: For each $l = 1, \ldots, L$ apply EMOS/NR or BMA to obtain a post-processed predictive CDF F_l.
2. Quantization: For each $l = 1, \ldots, L$ obtain a sample if size m from F_l, e.g. using

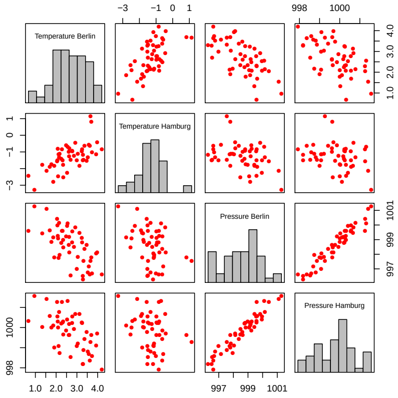
$$\tilde{x}_i = F_l^{-1}\left(\frac{i}{m+1}\right), \quad i = 1, \ldots, m$$

3. Ensemble reordering: Take C in Sklar's theorem to be the empirical copula of the raw ensemble, i.e., arrange the post-processed values in the same rank order as the raw ensemble values.

Schefzik, R., Thorarinsdottir, T. L., and Gneiting, T. (2013) **Uncertainty quantification in complex simulation models using ensemble copula coupling**. *Statistical Science*, 28, 616–640.

Example

24-hour ahead ECMWF ensemble forecast of surface temperature and pressure at Berlin and Hamburg valid November 8, 2011 before and after NR post-processing



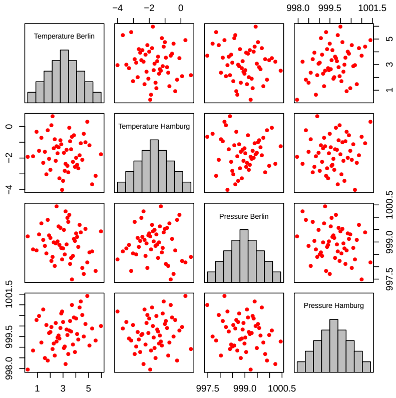


figure courtesy of Roman Schefzik

Example

24-hour ahead ECMWF ensemble forecast of surface temperature and pressure at Berlin and Hamburg valid November 8, 2011 before and after NR + ECC post-processing

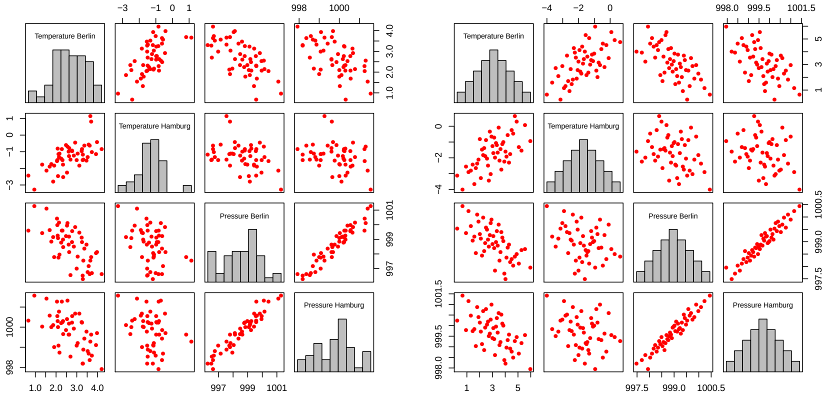


figure courtesy of Roman Schefzik

Overview

- Making and evaluating probabilistic forecasts
- Weather forecast ensembles and statistical post-processing
- Multivariate copula-based post-processing approaches
- Neural network approaches

Post-processing with neural network methods

Thus far, post-processing research has been focused on 'classical' statistical modeling and estimation approaches.

- Can AI/neural network methods be used in this context?
- Which approaches work well in which situations?
- What are advantages and disadvantages?

The results presented in the following are based on an ongoing joint project with Stephan Rasp (LMU Munich). See `https://github.com/slerch/ppnn` for code and data.

Post-processing and neural networks

[ ]

Input Ensemble predictions (which variables and summary statistics, for which locations/stations/grid points?)

Output Distribution parameters (or simulation draws? or interval probabilities?)

Data

All data are publicly available (around 30 GB after processing).

- observations from stations of the Deutscher Wetterdienst
  - around 500 stations in Germany
  - focus on temperature, more variables available
- ensemble forecasts (48h ahead) from the TIGGE archive
  - focus on ECMWF forecasts
  - temperature forecasts + 20–30 additional covariates chosen based on meteorological knowledge
  - available from 2008–2016
  - available on grid (0.5 degree resolution)
  - processing steps: interpolation to station locations, computation of summary statistics

Aim: Use data from 2015 to model temperatures in 2016.

Reminder: Standard NR post-processing

Temperature can be modeled using a Gaussian distribution,

$$y|x_1, \ldots, x_m \sim \mathcal{N}(\mu, \sigma^2),$$

where (in the most simple case)

$$\mu = a + b\,\bar{X} \quad \text{and} \quad \sigma^2 = c + d\,S^2,$$

with $x_1, \ldots, x_m$ denoting ensemble forecasts of temperature (i.e., no covariates used yet).

Reminder: Standard NR post-processing

Temperature can be modeled using a Gaussian distribution,

$$y|x_1, \ldots, x_m \sim \mathcal{N}(\mu, \sigma^2),$$

where (in the most simple case)

$$\mu = a + b\,\bar{X} \quad \text{and} \quad \sigma^2 = c + d\,S^2,$$

with $x_1, \ldots, x_m$ denoting ensemble forecasts of temperature (i.e., no covariates used yet).

Based on pairs of forecasts and observations from 2015, estimate model parameters a, b, c, d by minimizing the mean CRPS.

- choose local or regional training set

All results shown in the following are averages over all forecast cases in 2016.

Results

Model	Parameters	Mean CRPS
Raw ensemble		1.16
Traditional post-processing		
NR global	4	1.01
NR local	(4)	0.92
Network methods		

Adding covariate information via boosting

Adding additional variables to standard NR models is not straightforward as

- choice of suitable link function not obvious
- large number of additional parameters problematic for estimation

Here: Use boosting algorithm suggested by Messner et al. (2017):

Iterative optimization, updating only coefficients for covariates with largest effects (correlations) on partial derivatives, retain only those above a threshold.

Messner, J.W., Mayr, G.J. and Zeileis, A. (2017) **Nonhomogeneous boosting for predictor selection in ensemble postprocessing**. *Monthly Weather Review*, 145, 137–147.

Results

Model	Parameters	Mean CRPS
Raw ensemble		1.16
Traditional post-processing		
NR global	4	1.01
NR global + boosting	82	0.97
NR local	(4)	0.92
NR local + boosting	(82)	0.89
Network methods		

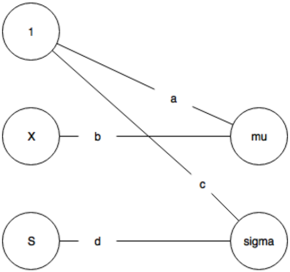
NR as a linear network

Recall that

$$y|x_1, \ldots, x_m \sim \mathcal{N}(\mu, \sigma^2),$$

where $\mu = a + b\,\bar{X}$ and $\sigma^2 = c + d\,S^2$.

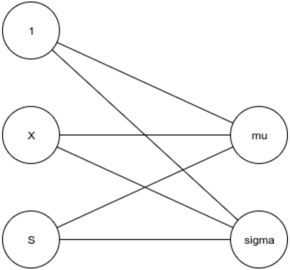
This can be viewed as a linear network:



- input: summary statistics of ensemble forecasts of temperature
- output: distribution parameters

Post-processing using a linear network

Using a fully connected linear network:



- weights and biases are determined by minimizing the mean CRPS in the training set as loss function
- should approximately re-produce results of NR global

Results

Model	Parameters	Mean CRPS
Raw ensemble		1.16
Traditional post-processing		
NR global	4	1.01
NR global + boosting	82	0.97
NR local	(4)	0.92
NR local + boosting	(82)	0.89
Network methods		
linear network	6	1.01

Extensions of linear networks

- auxiliary variables: Add (mean and standard deviation) of meteorological covariates as additional input features

Extensions of linear networks

- auxiliary variables: Add (mean and standard deviation) of meteorological covariates as additional input features
- local station information: Utilize local station information via embeddings, i.e., map station ID to n_{emb}-dimensional vector (rather than estimating separate models for all stations)

Extensions of linear networks

- auxiliary variables: Add (mean and standard deviation) of meteorological covariates as additional input features
- local station information: Utilize local station information via embeddings, i.e., map station ID to n_{emb}-dimensional vector (rather than estimating separate models for all stations)
- previous errors as features: Persistence of forecast errors in time motivates use of errors from *some previous* days as additional features

Results

Model	Parameters	Mean CRPS
Raw ensemble		1.16
Traditional post-processing		
NR global	4	1.01
NR global + boosting	82	0.97
NR local	(4)	0.92
NR local + boosting	(82)	0.89
Network methods		
linear network	6	1.01
linear network + aux. variables	82	0.92
linear network + embeddings	1048	0.91
linear network + fc. errors	126	0.89

Results

Model	Parameters	Mean CRPS
Raw ensemble		1.16
Traditional post-processing		
NR global	4	1.01
NR global + boosting	82	0.97
NR local	(4)	0.92
NR local + boosting	(82)	0.89
Network methods		
linear network	6	1.01
linear network + aux. variables	82	0.92
linear network + embeddings	1048	0.91
linear network + fc. errors	126	0.89
linear network + emb. + aux.var.	1160	0.88
linear network + emb. + aux.var. + fc.err.	1280	0.87

From linear to neural networks

Adding non-linearity via hidden layers (here: only one, with 500 nodes) results in additional improvements over the best traditional methods.

not much hyperparameter tuning thus far

Increased complexity results in worse predictions due to limited training data (overfitting!).

Results

Model	Parameters	Mean CRPS
Raw ensemble		1.16
Traditional post-processing		
NR global	4	1.01
NR global + boosting	82	0.97
NR local	(4)	0.92
NR local + boosting	(82)	0.89
Network methods		
linear network	6	1.01
linear network + aux. variables	82	0.92
linear network + embeddings	1048	0.91
linear network + fc. errors	126	0.89
linear network + emb. + aux.var.	1160	0.88
linear network + emb. + aux.var. + fc.err.	1280	0.87
neural network + emb. + aux.var.	3326	**0.83**

Things that didn't work

- recurrent neural networks: Include temporal information (sequence of previous forecasts) via recurrent units? However: No improvements, additional complexity results in overfitting.
- data augmentation techniques: In image recognition models, techniques such as randomly rotating, zooming or flipping pictures increase sample size and address overfitting. However, adding random noise does not lead to improvements here.

In general: positive effect of longer training periods (use of training data from 2008-2015 results in additional 6% improvement).

However: NWP model updates?

Discussion and outlook

- promising initial "proof of concept" results
- much faster than traditional methods
- particularly well suited for large data sets with many possible predictor variables
- (even more of a) *black-box* approach

- other, more complicated variables (wind, precipitation)?
- non-parametric interval-based variants?
- from station-based to grid-based models?
- What can we learn from networks about model deficiencies?

Questions you can – hopefully – answer now

- Why should predictions be probabilistic?
- How can probabilistic forecasts be evaluated, and what can mathematical statistics contribute?
- How are modern weather forecasts produced, and what is the role of statistics?
- Isn't this supposed to be an AI meetup?

Questions you can – hopefully – answer now

- Why should predictions be probabilistic?
- How can probabilistic forecasts be evaluated, and what can mathematical statistics contribute?
- How are modern weather forecasts produced, and what is the role of statistics?
- Isn't this supposed to be an AI meetup?

Thank you for your attention!