Statistical post-processing of ensemble forecasts: recent developments and current issues

Michael Scheuerer

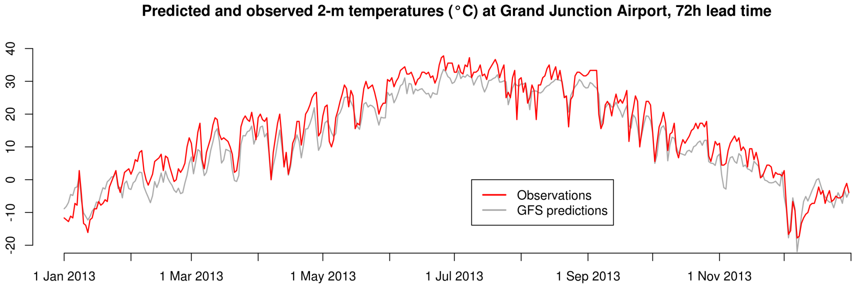
NOAA/ESRL, Physical Sciences Division

January 2016

Why statistical post-processing?

Despite continuous improvements to numerical weather prediction (NWP) systems, certain forecasts still suffer from systematic biases:

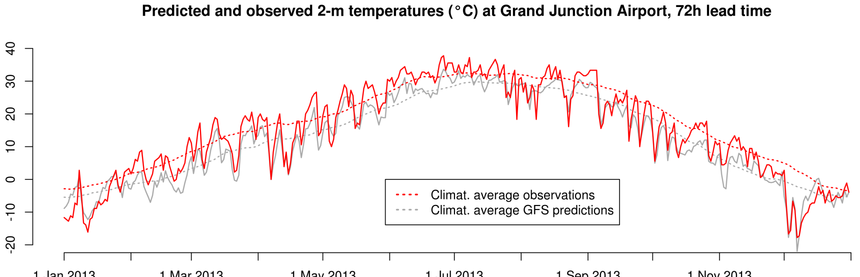
- insufficient model resolution
- less-than-optimal initial conditions
- etc.



Why statistical post-processing?

Despite continuous improvements to numerical weather prediction (NWP) systems, certain forecasts still suffer from systematic biases:

- insufficient model resolution
- less-than-optimal initial conditions
- etc.



Bias correction and MOS-type post-processing

If we have enough training data (past forecasts and observations) to estimate the respective climatological means $\mu_{cl,fcst}$ and $\mu_{cl,obs}$ for each day of the year, we can correct the systematic bias via

$$\tilde{x} = x - \mu_{cl,fcst} + \mu_{cl,obs}$$

Bias correction and MOS-type post-processing

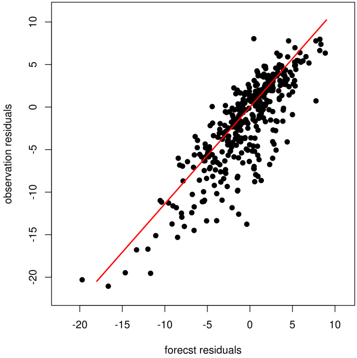
If we have enough training data (past forecasts and observations) to estimate the respective climatological means $\mu_{cl,fcst}$ and $\mu_{cl,obs}$ for each day of the year, we can correct the systematic bias via

$$\tilde{x} = x - \mu_{cl,fcst} + \mu_{cl,obs}$$

Or, we can go one step further and fit a regression model to forecasts and observations:

$$\tilde{x} = \mu_{cl,obs} + a \cdot (x - \mu_{cl,fcst}),$$

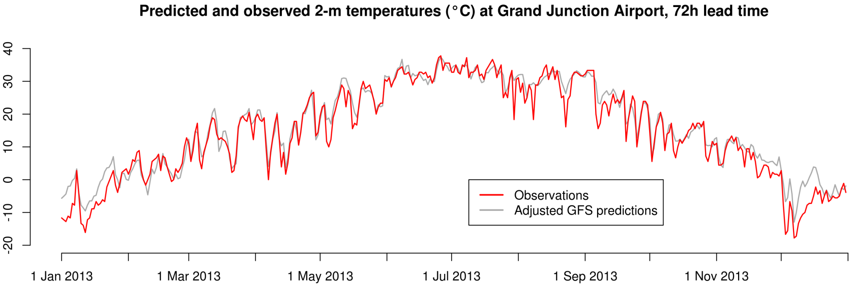
thus accounting also for forecast skill and obtaining an adjusted forecast $\tilde{x}$.



Bias correction and MOS-type post-processing

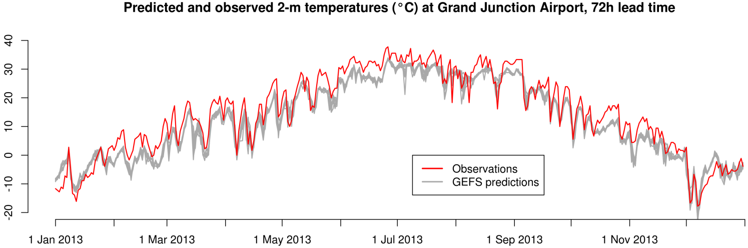
These regression-type ("MOS") adjustments

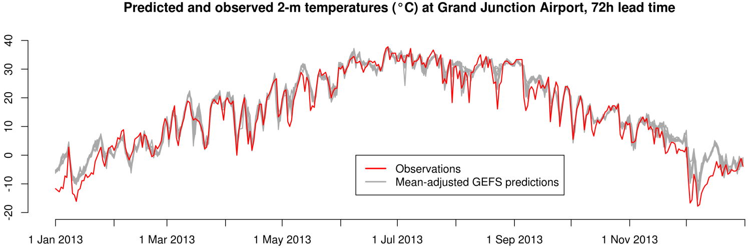
- result in mean squared error optimal forecasts
- can be extended to include additional predictors
- can be adapted to weather variables that require certain restrictions (non-negativity, etc.)



Ensemble post-processing

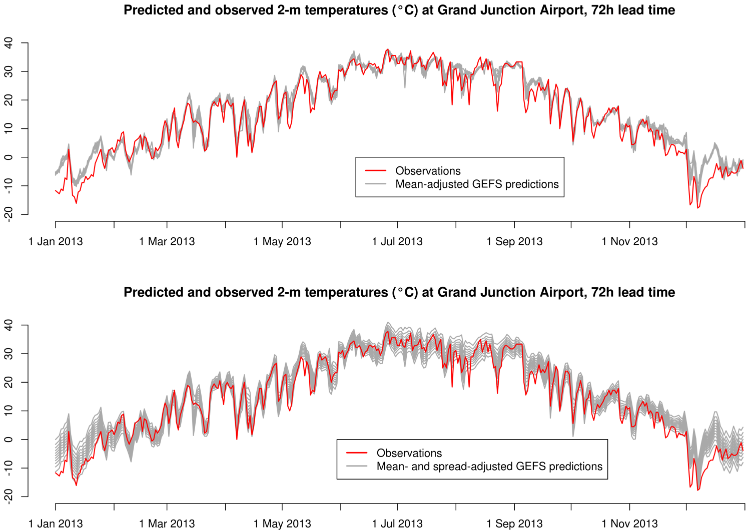
The same kind of correction can also be applied to ensemble forecasts:





Ensemble post-processing

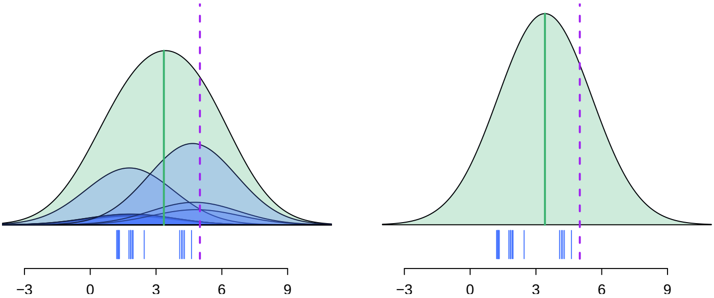
For ensembles, the spread needs to be adjusted in addition to the mean:



Ensemble post-processing

Approaches to ensemble post-processing:

- Kernel dressing methods, Bayesian Model Averaging (BMA)
- Nonhomogeneous Gaussian Regression (NGR, “EMOS”)
- Bayesian processor of Ensemble
- Similarity-based (“analog”) techniques
- Member-by-member approaches
- etc.



Probability forecasts

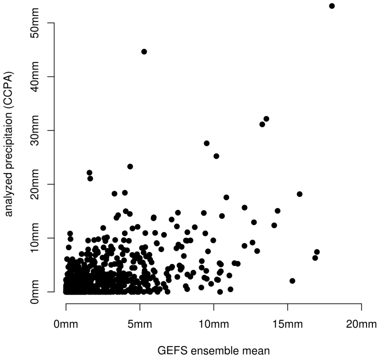
Probability forecasts for events (e.g. "rainfall amounts exceed 10mm") can be derived from calibrated ensemble forecasts or predictive distributions.

Or, event probabilities can be modeled directly, e.g. via logistic regression:

$$\text{logit}\big(P(y > 10mm)\big) = \beta_0 + \beta_1 \cdot x$$

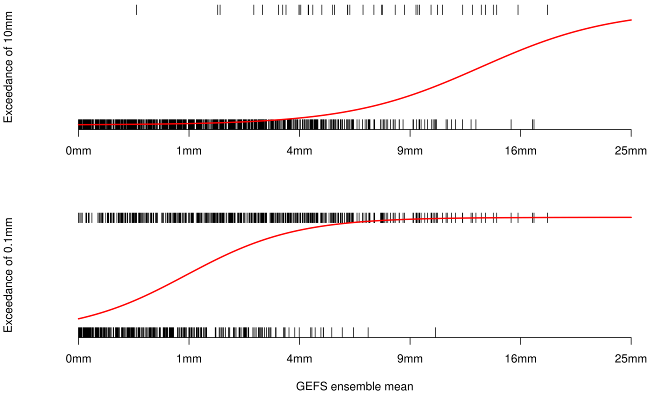
Example:

60 to 72-h Precipitation accumulations over Seattle during the winter season.



Logistic regression and extended logistic regression

Logistic regression fits a separate model for each threshold:

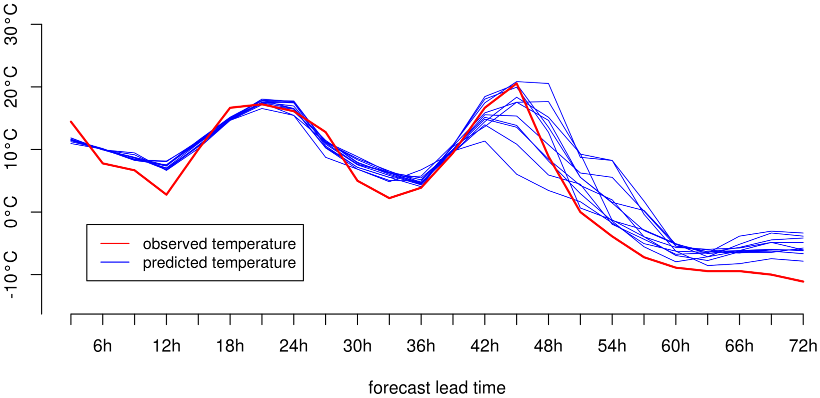


Extended logistic regression links the different threshold probabilities and estimates a joint model, thus yielding again a full predictive distribution.

Multivariate post-processing

Consider a probabilistic forecast of a multivariate quantity, where multivariate may refer to different variables, or the same variable at different time points and/or locations in space.

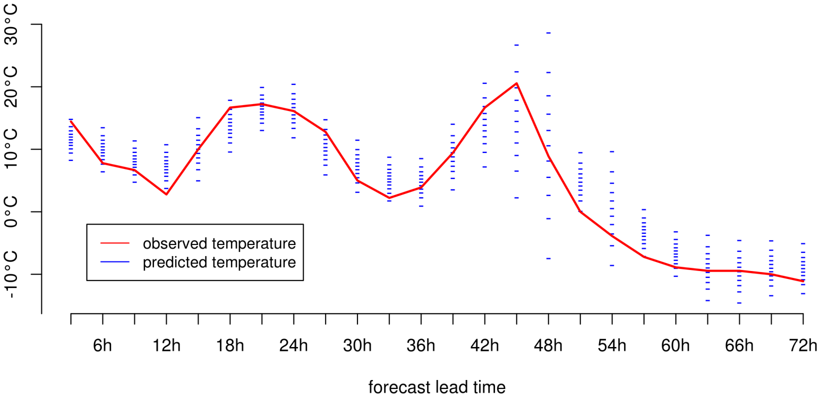
Example: Temperature forecasts at Denver for lead times up to 72-h



Multivariate post-processing

Applying the post-processing techniques discussed above yields calibrated forecasts at each lead time separately. How can we re-create forecast trajectories with adequate temporal correlations?

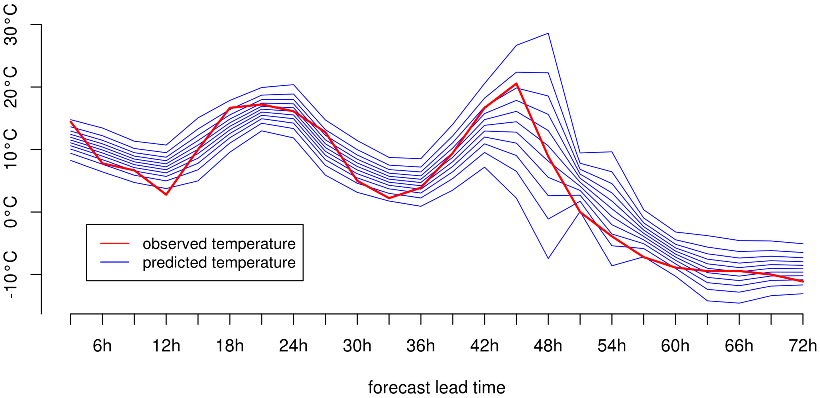
Example: Temperature forecasts at Denver for lead times up to 72-h



Multivariate post-processing

Applying the post-processing techniques discussed above yields calibrated forecasts at each lead time separately. How can we re-create forecast trajectories with adequate temporal correlations?

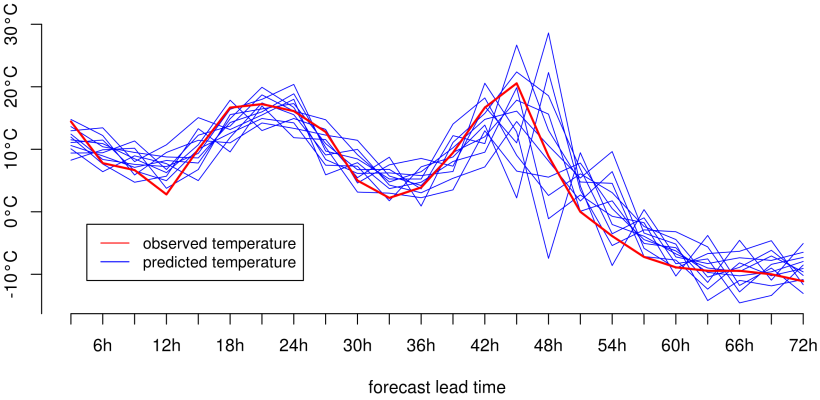
Example: Temperature forecasts at Denver for lead times up to 72-h



Multivariate post-processing

Applying the post-processing techniques discussed above yields calibrated forecasts at each lead time separately. How can we re-create forecast trajectories with adequate temporal correlations?

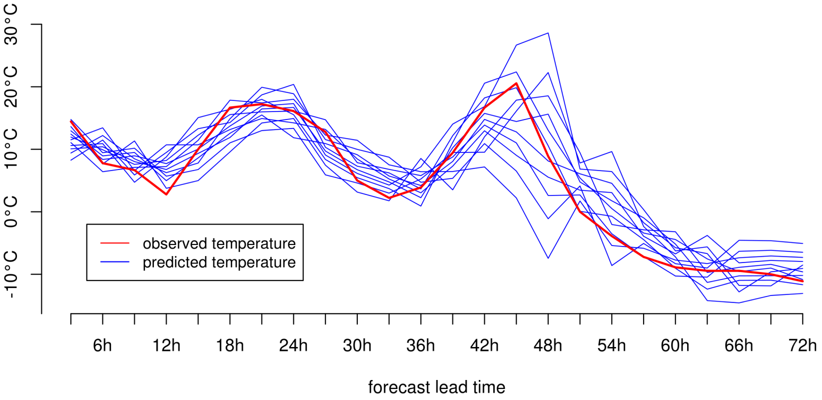
Example: Temperature forecasts at Denver for lead times up to 72-h



Multivariate post-processing

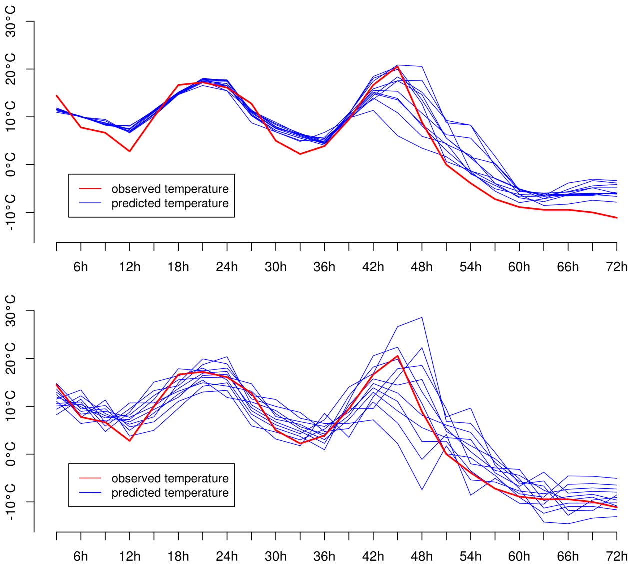
Applying the post-processing techniques discussed above yields calibrated forecasts at each lead time separately. How can we re-create forecast trajectories with adequate temporal correlations?

Example: Temperature forecasts at Denver for lead times up to 72-h



Using multivariate information from raw ensemble forecasts

Ensemble copula coupling (ECC):



Idea: retain the ordering (and thus the rank correlations) of the raw ensemble forecasts but replace their values by those derived from the calibrated marginal distributions.

Special case: member-by-member calibration

Using multivariate information from past observations

Schaake Shuffle:

Proceed as with ECC, but use the rank order of *past obervations at the same or similar days of the year* instead of the ranks of today's ensemble forecasts.

Similarity-based Schaake Shuffle:

Use again observation ranks but select the historic dates based on similarity of the respective forecasts.

Statistical dependence models:

Fit a statistical dependence model (Gaussian copulas, Gaussian random fields) using forecast error statistics at historic dates.

Two main approaches for multivariate post-processing

1. Use multivariate information from raw ensemble forecasts

+ flow-dependent, physics-based correlations
+ potentially different correlations for different forecast magnitudes

– spurious correlations in the raw ensemble may be amplified
– multivariate features that are not resolved by the NWP model are not accounted for
– ensemble size limits the representativeness of multivariate features

Two main approaches for multivariate post-processing

1. Use multivariate information from raw ensemble forecasts

   + flow-dependent, physics-based correlations
   + potentially different correlations for different forecast magnitudes

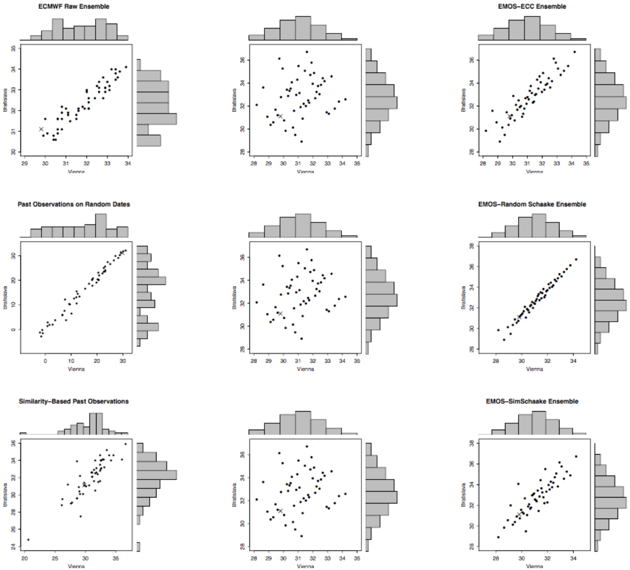
   – spurious correlations in the raw ensemble may be amplified
   – multivariate features that are not resolved by the NWP model are not accounted for
   – ensemble size limits the representativeness of multivariate features

2. Use multivariate information from past observations

   + more realistic error structures
   + downscaling of dependence information

   – multivariate information is not flow-dependent
   – extra efforts are required to model correlations that depend on the forecast magnitude

Bivariate example: ECC vs. Schaake vs. SimSchaake

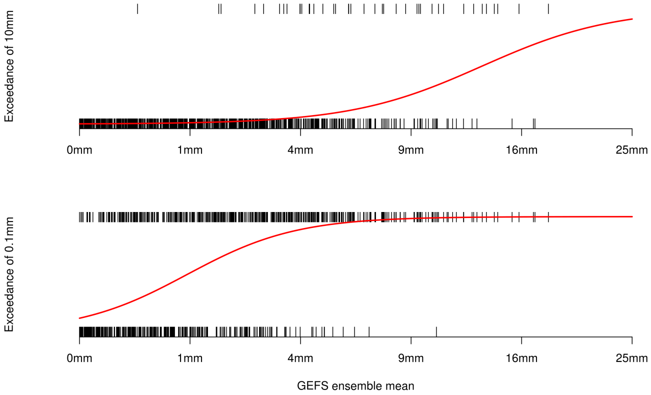


24 hour ahead EMOS-calibrated temperature forecasts (in °C) at Vienna and Bratislava valid on 9 July 2011, 1200 UTC.

Image courtesy of Roman Schefzik.

Probabilistic forecasts of rare events

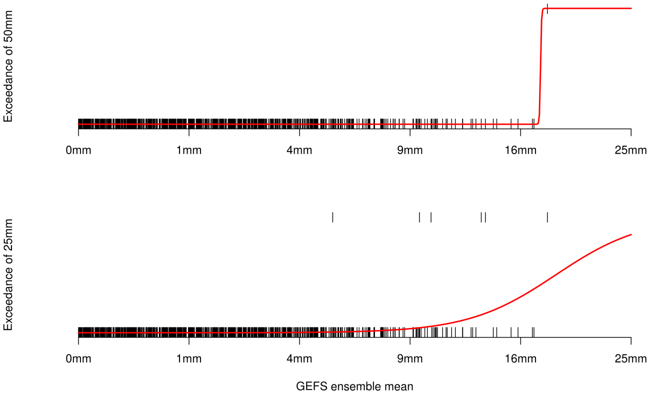
Fitting a logistic regression model for high thresholds becomes increasingly difficult:



Parametric assumptions can mitigate the problems that come with modeling rare events, but limited training sample size remains a concern.

Probabilistic forecasts of rare events

Fitting a logistic regression model for high thresholds becomes increasingly difficult:



Parametric assumptions can mitigate the problems that come with modeling rare events, but limited training sample size remains a concern.

Options for getting a sufficiently large training sample

1. Reforecasts!
   - \+ no compromises, no biases
   - \+ ideally cover several years, thus variations in climatology
   - − expensive

2. Regional post-processing, supplemental locations, random field models that link locations statistically
   - \+ can reduce the need for reforecasts
   - − linking/combining less than perfectly similar locations entails biases

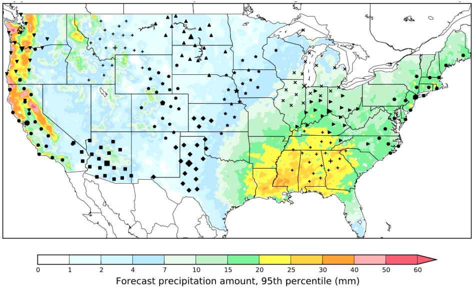
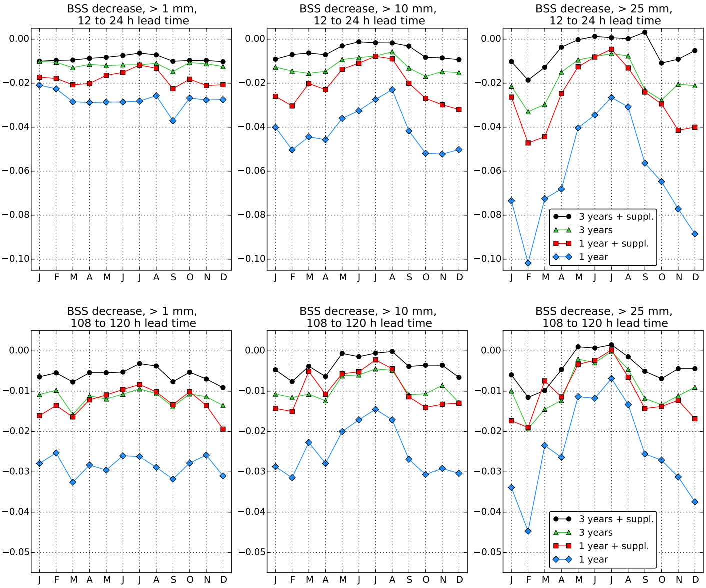


Image courtesy of Tom Hamill

Example: Loss of skill relative to a 11-year training sample



Brier skill scores for an EMOS-type post-processing method for precipitation amounts.

Rare event verification / Guide to immoral post-processing

Public and media attention usually focus on predictive performance for the subset of cases where some *high impact event* has happened, e.g.

"Bad data failed to predict Nashville Flood"

NBC, 2011

Clearly, these cases are of higher public interest than more 'ordinary' events. Scientifically, however, a verification strategy for probabilistic forecasts of the form

- select the cases where the outcome was extreme
- discard all non-extreme cases
- proceed with the evaluation using standard proper scoring rules

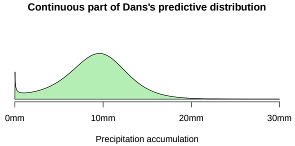
is very problematic!

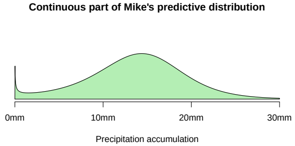
It discourages honest forecasting and encourages exaggeration.

Example: Bad verification rewards cheaters

Consider again the 60 to 72-h precipitation forecast dataset for cool season precipitation over Seattle. We use cross-validation (leave out one of the 12 years of data at a time) and compare two forecasters:

1. Dan: calibrates the GEFS forecasts via extended logistic regression
2. Mike: scales Dan's predictive distributions by a factor 1.5



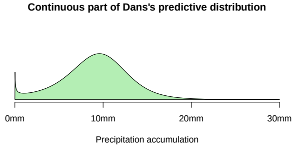


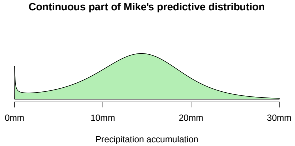
	all cases
CRPS: Dan	**1.15**
CRPS: Mike	1.24

Example: Bad verification rewards cheaters

Consider again the 60 to 72-h precipitation forecast dataset for cool season precipitation over Seattle. We use cross-validation (leave out one of the 12 years of data at a time) and compare two forecasters:

1. Dan: calibrates the GEFS forecasts via extended logistic regression
2. Mike: scales Dan's predictive distributions by a factor 1.5





	all cases	$y > 0.1$	$y > 5$	$y > 10$	$y > 25$
CRPS: Dan	**1.15**	**2.31**	5.02	8.70	19.5
CRPS: Mike	1.24	2.41	**4.51**	**6.98**	**13.5**

Literature I

Glahn, H.R., and Lowry, D.A.
The use of Model Output Statistics (MOS) in objective weather forecasting.
J. Appl. Meteor., 11:1203–1211, 1972.

Carter, G.M., Dallavalle, J.P., and Glahn, H.R.
Statistical forecasts based on the National Meteorological Center's numerical weather prediction system.
Wea. Forecasting, 4:401–412, 1989.

Roulston, M.S., and Smith, L.A.
Combining dynamical and statistical ensembles.
Tellus, 55A:16–30, 2003.

Wang, X., and Bishop, C.H.
Improvement of ensemble reliability with a new dressing kernel.
Quart. J. Roy. Meteor. Soc., 131:965–986, 2004.

Raftery, A.E., Gneiting, T., Balabdaoui, F., and Polakowski, M.
Using Bayesian model averaging to calibrate forecast ensembles.
Mon. Wea. Rev., 133:1155–1174, 2005.

Gneiting, T., Raftery, A.E., Westveld, A.H., and Goldman, T.
Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation.
Mon. Wea. Rev.,133:1098–1118, 2005.

Literature II

Krzysztofowicz, R.
Bayesian Processor of Ensemble: concept and development.
Proc. 19th Conf. Probability and Statistics in the Atmospheric Sciences, vol. 4.5. Seattle: American Meteorological Society, 2008.

Hamill, T.M., and Whitaker, J.S.
Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application.
Mon. Wea. Rev., 134:3209–3229, 2006.

Van Schaeybroeck, B., and Vannitsem, S.
Ensemble post-processing using member-by-member approaches.
Quart. J. Roy. Meteor. Soc., 141:807–818, 2015.

Wilks, D.S.
Extending logistic regression to provide full-probability-distribution MOS forecasts.
Meteor. Applic., 16:361–368, 2009.

Schefzik, R., Thorarinsdottir, T.L., and Gneiting, T.
Uncertainty quantification in complex simulation models using ensemble copula coupling.
Stat. Sci., 28:616–640, 2013.

Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.
The Schaake shuffle: A method for reconstructing spaceâtime variability in forecasted precipitation and temperature fields.
J. Hydrometeor., 5:243–262, 2004.

Literature III

Schefzik, R.
A similarity-based implementation of the Schaake shuffle.
preprint, http://arxiv.org/pdf/1507.02079.pdf

Hamill, T.M., Scheuerer, M. and Bates, G.T.
Analog probabilistic precipitation forecasts using GEFS Reforecasts and Climatology-Calibrated Precipitation Analyses.
Mon. Wea. Rev., 143:3300–3309, 2015.

Möller, A., Thorarinsdottir, T.L., Lenkoski, A., and Gneiting, T.
Spatially adaptive, Bayesian estimation for probabilistic temperature forecasts.
preprint, http://arxiv.org/pdf/1507.05066.pdf

Scheuerer, M. and Hamill, T.M.
Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions.
Mon. Wea. Rev., 143:4578–4596, 2015.

Lerch, S., Thorarinsdottir, T.L., Ravazzolo, F., and Gneiting, T.
Forecaster's dilemma: Extreme events and forecast evaluation.
preprint, http://arxiv.org/pdf/1512.09244v1.pdf