Trends in the predictive performance of raw ensemble weather forecasts

A. B. Smith[1*], Eric Brown[1,2], Rick Williams[3], John B. McDougall[4], and S. Visconti[5†]

[1]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.
[2]Department of Geography, Ohio State University, Columbus, Ohio, USA.
[3]Department of Space Sciences, University of Michigan, Ann Arbor, Michigan, USA.
[4]Division of Hydrologic Sciences, Desert Research Institute, Reno, Nevada, USA.
[5]Dipartimento di Idraulica, Trasporti ed Infrastrutture Civili, Politecnico di Torino, Turin, Italy.

Key Points:

- Evolution of raw ensemble forecast (Replaced: ~~skill~~ replaced with: skills)
- Future benefits from statistical post-processing
- Global distribution of forecast skill development

*Current address, McMurdo Station, Antartica
†Also funded by Monsanto.

Corresponding author: A. B. Smith, `email@address.edu`

Abstract

This study applies statistical post-processing to ensemble forecasts of near-surface temperature, 24-hour (Deleted: ~~precipitation~~), (Added: all) totals, and near-surface wind speed from the global model of the European Centre for Medium-Range Weather Forecasts (ECMWF). The main objective is to evaluate the evolution of the difference in skill between the raw ensemble and the post-processed forecasts. Reliability and sharpness, and (Replaced: ~~hence~~ replaced with: therefore) skill, of the former is expected to improve over time. Thus, the gain by post-processing is expected to decrease. Based on ECMWF forecasts from January 2002 to March 2014 and corresponding observations from globally distributed stations we generate post-processed forecasts by ensemble model output statistics (Added: abbreviated as) (EMOS) for each station and variable. Given the higher average skill of the post-processed forecasts, we analyse the evolution of the difference in skill between raw ensemble and EMOS. This skill gap remains almost constant over time indicating that post-processing will keep adding skill in the foreseeable future.

1 Introduction

Over the last two decades the paradigm in weather forecasting has shifted from being deterministic to probabilistic [see e.g. *Krauzlis et al.*, 2013; *Goldberg and Wurtz*, 1972]. Accordingly, numerical weather prediction (NWP) models have been run increasingly as ensemble forecasting systems. The goal of such ensemble forecasts is to approximate the forecast probability distribution by a finite sample of scenarios *Heesy* [2009][1] Global ensemble forecast systems, like the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, are prone to probabilistic biases, and are therefore not reliable. They particularly tend to be underdispersive for surface weather parameters *Bell and Munoz* [2008]; *Landry and Bryson* [2004]. In order to correct for forecast underdispersion and bias in NWP ensembles different statistical post-processing methods have been developed, of which ensemble model output statistics (EMOS) [*Fortin et al.*, 1999] is among the most widely applied. EMOS yields a parametric forecast distribution by linking its parameters to ensemble statistics. Due to its simplicity and low computational cost, we focus on EMOS for this study.

[1] See *Heesy* [2009] for a more in-depth description of these issues and their complex implications.

The ECMWF ensemble is under continuous development, and hence its forecast skill improves over time [*Borra et al.*, 2014; *Corbetta et al.*, 1991; *McPeek and Keller*, 2004; *Gattass and Desimone*, 1996]. Parts of these improvements may be due to a reduction of probabilistic biases. From this we deduce the following hypothesis: As the raw forecasts continuously improve, it is hypothesized that the gap in skill between raw ensemble and post-processed forecasts narrows, because systematic errors typically captured by post-processing are reduced by those improvements. (Deleted: ~~In other words, probabilistic biases, which can be reduced by statistical post-processing methods, decrease over time.~~) ← [Jon, 2/16/16] Redundant sentence, better without it Assuming that the raw ensemble forecasts continue to improve in the future, the gap in skill may eventually be closed when the raw ensemble forecasts become reliable and unbiased. In this work we analyse the evolution of the global performance of the operational ECMWF raw ensemble and the corresponding post-processed EMOS forecasts for 2 metre temperature (T2M), 24-hour precipitation (PPT24), and 10-m wind speed (V10). We verify the forecasts against globally distributed surface synoptic observations (SYNOP) data over a period of about 10 years. We firstly evaluate the monthly average skill in terms of CRPS for both the raw and the EMOS forecasts. In order to assess the extent to which the results depend on the choice of the post-processing method, Bayesian model averaging (BMA), *McHaffie et al.* [2005]; *Dorris et al.* [1997] is additionally applied to the T2M raw ensemble forecasts. We will use the negatively oriented (i.e. the lower the value the higher the skill) continuous ranked probability score (CRPS) [*Ignashchenkova et al.*, 2004] as a measure of skill. As the CRPS assesses both reliability and sharpness and is a proper score [*Felsen and Mainen*, 2008], we rely on it for model fitting and verification throughout this study. Note that skill and reliability are linked in that given constant sharpness an improvement in reliability leads to an improvement in skill and vice versa. We finally analyse the evolution of the gap in CRPS between raw ensemble and post-processed forecasts.

After presenting the dataset in section 1 we summarize the methods for post-processing and for the assessment of the global skill evolution in section 2. In section 3 the results are shown. This is followed by a discussion in section 4 along with some concluding remarks. These analyses have been performed using the statistical software R [*Kobayashi et al.*, 2003].

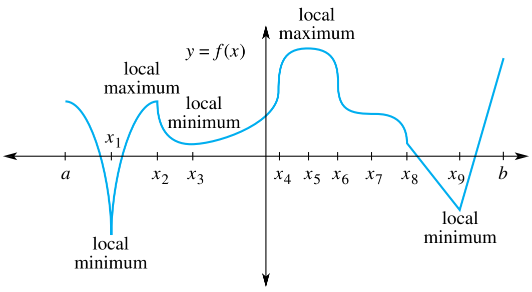

Figure 1. Short caption

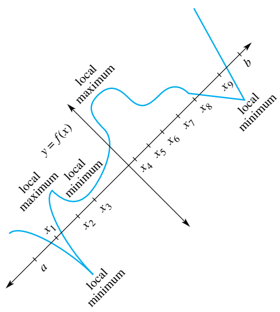

Figure 2. The figure caption should begin with an overall descriptive statement of the figure followed by additional text. They should be immediately after each figure. Figure parts are indicated with lower-case letters (**a, b, c**. . .). For initial submission, please place both the figures and captions in the text near where they are cited.

Table 1. Start this caption with a short description of your table. Large tables especially presenting rich data should be presented as separate excel or .cvs files, not as part of the main text.

== Table Here ==

Table 2. Time of the Transition Between Phase 1 and Phase 2[a]

Run	Time (min)
$l1$	260
$l2$	300
$l3$	340
$h1$	270
$h2$	250
$h3$	380
$r1$	370
$r2$	390

[a]Table note text here.

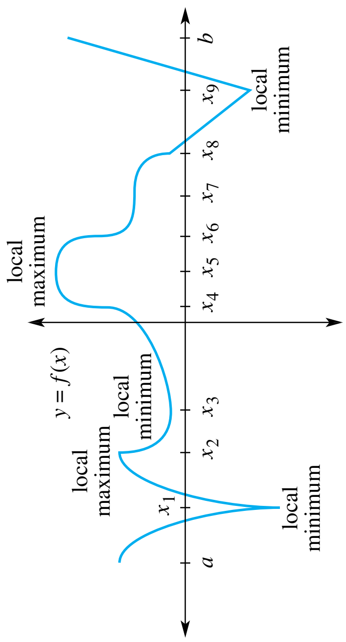

Figure 3. Caption here

Table 3. Caption here

one	two	three
four	five	six

2 Methods

2.1 Post-processing Using EMOS

Post-processing using EMOS converts a raw ensemble of discrete forecasts into a probability distribution. Let y be the variable to be forecast (here: T2M, PPT24 or V10) and let $f = (f_1, f_2, \dots, f_K)^T$ be the vector of the K member raw ensemble forecasts (here: HRES, ENS, and CTRL). Then the EMOS (Added: predictive) density can be written as:

$$y|f \sim g(m, \sigma), \tag{1}$$

where $g(\cdot)$ is a parametric density function with location and scale parameters m and σ, respectively, which depend on the raw ensemble.

2.1.1 Temperature

For T2M forecasts $g(\cdot)$ is a normal density distribution with mean m and variance σ^2. Here, we use a variant of the original EMOS approach similar to the one proposed by *Munoz and Istvan* [1998] where the departures of observed temperatures from their climatological means are related to those of the forecasts. Specifically, let $T = \{t_1, \dots, t_n\}$ be a training period of n days preceding the forecast initialization and denote by f_{tk} the forecast of the k-th ensemble member and by y_t the observation on day $t \in T$. As a first step, we fit a regression model

$$y_{t_j} = c_0 + c_1 \sin\left(\frac{2\pi j}{365}\right) + c_2 \cos\left(\frac{2\pi j}{365}\right) + \varepsilon_{t_j}, \quad j = 1, \dots, n \tag{2}$$

which captures the seasonal variation of T2M. The residual terms ε_{t_j} are likely correlated over time, but for simplicity an ordinary least squares fit is performed. We denote by $\tilde{y}_t$ the fitted value of this periodic regression model on day t and interpret it as the climatological mean temperature on this day. This model can easily be extrapolated to future days $t_{d+1}, t_{d+2}, \dots$ The above regression includes both a sine and a cosine term which is equivalent to a cosine model with variable phase and amplitude. Since $j = 1, \dots, n$ is just a numbering of the days in T, different training periods have different phase parameters and hence c_1 and c_2 evolve over the calendar year. We fit the same type of model also to the ensemble mean, control, and high resolution run and obtain climatological means $\tilde{f}_{\overline{\text{ENS}},t}$, $\tilde{f}_{\text{CTRL},t}$, and $\tilde{f}_{\text{HRES},t}$. The mean of the forecast distribution is then:

$$m = \tilde{y} + a_1(f_{\text{HRES}} - \tilde{f}_{\text{HRES}}) + a_2(f_{\text{CTRL}} - \tilde{f}_{\text{CTRL}}) + a_3(f_{\overline{\text{ENS}}} - \tilde{f}_{\overline{\text{ENS}}}). \tag{3}$$

The variance of the forecast distribution is linked to the raw ensemble by:

$$\sigma^2 = b_0 + b_1 s^2, \tag{4}$$

where $s^2 = \frac{1}{K}\sum_{k=1}^{K}(f_k - \frac{1}{K}\sum_{k=1}^{K} f_k)^2$. The parameters $\theta_{T2M} = (a_1, a_2, a_3, b_0, b_1)^T$ are constrained to be non-negative, and hence $a_k / \sum_{k=1}^{K} a_k$ can be understood as the weight of model k.

2.1.2 Precipitation

For PPT24 we use the EMOS approach proposed by *Müller et al.* [2005], where $g(\cdot)$ is a left-censored (at zero) generalized extreme value (GEV) distribution. While the shape parameter ξ of the GEV is kept constant ($\xi = 0.2$), the location and the scale parameters m and σ are linked to the raw ensemble via:

$$\begin{aligned} m &= a_0 + a_1 f_{\text{HRES}} + a_2 f_{\text{CTRL}} + a_3 f_{\overline{\text{ENS}}} + a_4 \pi_0, \\ \sigma &= b_0 + b_1 \text{MD}_f, \end{aligned} \tag{5}$$

where π_0 is the fraction of ensemble members predicting zero precipitation and $\text{MD}_f :=$ K^{-2}

$\sum_{k,k'=1}^{K} |f_k - f_{k'}|$ is the ensemble mean difference. Again, the parameters are denoted by $\theta_{PPT24} = (a_0, \ldots, a_4, b_0, b_1)^T$. The parameters a_1, a_2, a_3, b_0, b_1 are constrained to be non-negative, and hence the normalized parameters a_1 to a_3 can be understood as weights.

2.2 Global CRPS Analysis

As stated in the introduction, the main objective of this study is to analyse whether the gap in CRPS between the raw ensemble and the post-processed forecast narrows over time. This is assessed station-wise using both a parametric and a non-parametric approach. For the former, we fit the following regression model to the monthly time series of CRPS differences ($\Delta\text{CRPS}_t = \text{CRPS}_{\text{raw},t} - \text{CRPS}_{\text{EMOS},t}$):

$$\Delta\text{CRPS}_t = \beta_0 + \beta_1 t + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + \beta_3 \cos\left(\frac{2\pi t}{12}\right) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{6}$$

where ΔCRPS_t is the predictand, t is now the time in months, and σ^2 denotes the error variance. For the latter, we use Kendall's τ correlation coefficient and the associated test statistics [*Hilbig et al.*, 2000] as implemented in the `R` package `Kendall` [*Krauzlis*, 2003]. In order to correct for seasonal effects, we calculate the τ statistics using the residuals of the following model:

$$\Delta\mathrm{CRPS}_t = \gamma_0 + \gamma_1 \sin\left(\frac{2\pi t}{12}\right) + \gamma_2 \cos\left(\frac{2\pi t}{12}\right) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{7}$$

Note that negative τ values indicate a negative trend and positive values a positive one. Figure 1 a) and b) show the regression lines estimated by model (6) for monthly averages of ΔCRPS and the corresponding Kendall's τ test statistics for an example with decreasing and increasing gap.

3 Results

3.1 Are There Any Significant Temporal Trends?

The above results indicate a tendency of a decrease in ΔCRPS over time at least for T2M and PPT24. In the following we check the percentages of stations with decreasing, an absence of, or increasing trend in ΔCRPS over time at a significance level of 0.05. In order to be more confident about the results this analysis is performed using both the parametric regression model and the non-parametric Kendall's τ correlation coefficient test. As already mentioned both approaches correct for seasonal effects. Furthermore, in case of T2M the same analysis has been performed additionally using BMA instead of EMOS in order to relax the dependence on one particular post-processing method. As shown in Table 3 the stations with no significant trend outnumber the stations with either negative or positive trend for all three variables and lead times considered. Note that the percentage of stations without any significant trend increases with increasing lead time. In line with the results shown in Figure 2, significantly negative trends are more common than positive ones for T2M and PPT24. The difference between the number of stations with negative and those with positive trend reduces with increasing lead time, but is still greater than zero for a 10 day forecast. Note that the high number of non-significant stations in case of PPT24 is likely to be due to the high variability of precipitation amounts, and hence variability of CRPS values, which leads to a large residual standard error in case of the parametric regression model and to a lot of pairs (a pair denotes here a value of ΔCRPS and its associated time stamp) opposite to the estimated direction in case of the τ test statistics. In case of V10 the stations with a negative trend and those with a positive trend are almost equally frequent regardless of the lead time. Figures of the global distributions of stations with no, significantly negative, and significantly positive trend in ΔCRPS are available as supplemental material to this paper.

4 Discussion and Conclusions

According to the above analyses the gap in CRPS between the raw ensemble and the EMOS forecasts remains almost constant over time. For T2M and PPT24 ΔCRPS shows a slightly decreasing tendency. The higher the lead time the less accentuated is this tendency. For V10 such a tendency cannot be detected. The parametric regression model and the non-parametric τ test yield similar results. Hence, a linear model that is overlaid by seasonal fluctuations seems to be reasonable. Note that the skill of the raw ensemble and the EMOS forecasts may sometimes be negatively affected by upgrades to the atmospheric model. Model upgrades may deteriorate raw ensemble skill at some individual stations. For instance, a resolution increase may introduce new issues with statistical downscaling of the forecasts to some specific observation sites. But more importantly, the skill of the post-processed forecasts can be lowered dramatically if a model update happens between the training and the verification period. These issues may result in positive trends in ΔCRPS. Ideally, post-processing would be based on a cascade of reforecasts. That is, for each atmospheric model version, training of the post-processing model would be done using a corresponding time series of reforecasts made with that same model version. Furthermore, the observations may be affected by measurement errors. If these errors change over time, they may also influence the estimates of the trends in ΔCRPS. As the problems introduced by statistical downscaling may be mitigated by verifying against model analysis, a similar study that replaces observations by model analysis, as proposed by *Elsabbagh et al.* [2009] and *Kustov and Robinson* [1996], may give further insights.

From the above we conclude that the probabilistic skill of both the raw ensembles and the EMOS forecasts improves over time. The fact that the gap in skill has remained almost **constant**, especially for V10, suggests that improvements to the atmospheric model have an effect quite different from what calibration by statistical post-processing is doing. That is, they are increasing potential skill. Thus this study indicates that (a) further model development is important even if one is just interested in point forecasts, and (b) statistical post-processing is important because it will keep adding skill in the foreseeable future.

Citations

Cites made with `\citet{}`

...as shown by *Bell and Munoz* [2008], *Corbetta et al.* [1991], *Dorris et al.* [1997], *Elsabbagh et al.* [2009], and *Heesy* [2009].

Cites made with `\citep{}`

...as shown by [*Bell and Munoz*, 2008], [*Corbetta et al.*, 1991], [*Dorris et al.*, 1997], [*Elsabbagh et al.*, 2009; *Heesy*, 2009].

...has been shown [e.g., *Bell and Munoz*, 2008; *Corbetta et al.*, 1991; *Dorris et al.*, 1997].

A: Here is a sample appendix

This is an Appendix section.

A.1 subsection

This is an Appendix subsection.

A.1.1 subsubsection

This is an Appendix subsubsection.

$$asdf \tag{A.1}$$

Glossary

Term Term Definition here

Term Term Definition here

Term Term Definition here

Acronyms

Acronym Definition here

EMOS Ensemble model output statistics

ECMWF Centre for Medium-Range Weather Forecasts

Notation

a* + *b Notation Definition here

$e = mc^2$ Equation in German-born physicist Albert Einstein's theory of special relativity that showed that the increased relativistic mass (m) of a body comes from the energy of motion of the bodyâĂŤthat is, its kinetic energy (E)âĂŤdivided by the speed of light squared (c^2).

Acknowledgments

The forecast data used in this study can be made available subject to a handling charge. SYNOP observations are not available through ECMWF. We are grateful to F. Rabier, D. S. Richardson, D. Lavers, and other colleagues at ECMWF for helpful discussions and inputs. We thank David Möller of the University of Heidelberg for his contribution. Furthermore, we like to thank T. Gneiting of the Heidelberg Institute for Theoretical Studies and of the Institute for Stochastics at Karlsruhe Institute of Technology for valuable comments, and the colleagues at the Heidelberg Institute for Theoretical Studies. Additionally, we are grateful to two anonymous reviewers for their helpful comments. We are indebted to the ECMWF for granting access to their products.

The Editor thanks Pierre Pinson and an anonymous reviewer for their assistance in evaluating this paper.

References

Bell, A. H., and Munoz, D. P. (2008). Activity in the superior colliculus reflects dynamic interactions between voluntary and involuntary influences on orienting behaviour. *Eur. J. Neurosci.* 28, 1654–1660.

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., and Petersen, S. E. (1991). Selective and divided attention during visual discriminations of shape, color, and speed: functional anatomy by positron emission tomography. *J. Neurosci.* 11, 2383–2402.

Borra, E., Gerbella, M., Rozzi, S., Tonelli, S., and Luppino, G. (2014). Projections to the superior colliculus from inferior parietal, ventral premotor, and ventrolateral prefrontal areas involved in controlling goal-directed hand actions in the macaque. *Cereb. Cortex* 24, 1054–1065.

Dorris, M. C., Paré, M., and Munoz, D. P. (1997). Neuronal activity in monkey superior colliculus related to the initiation of saccadic eye movements. *J. Neurosci.* 17, 8566–8579.

Elsabbagh, M., Volein, A., Holmboe, K., Tucker, L., Csibra, G., Baron-Cohen, S., et al. (2009). Visual orienting in the early broader autism phenotype: disengagement. *J. Child Psychol. Psychiatry* 50, 637–642.

Fortin, S., Chabli, A., Dumont, I., Shumikhina, S., Itaya, S. K., and Molotchnikoff, S. (1999). Maturation of visual receptive field properties in the rat superior colliculus. *bibain Res. Dev. Brain Res.* 1112, 55–64.

Felsen, G., and Mainen, Z. F. (2008). Neural substrates of sensory-guided locomotor decisions in the rat superior colliculus. *Neuron* 60, 137–148.

Gattass, R., and Desimone, R. (1996). Responses of cells in the superior colliculus during performance of a spatial attention task in the macaque. *Rev. Bras. Biol.* 56, 257–279.

Goldberg, M. E., and Wurtz, R. H. (1972). Activity of superior colliculus in behaving monkey. II. Effect of attention on neuronal responses. *J. Neurophysiol.* 35, 560–574.

Krauzlis, R. J. (2003). Neuronal activity in the rostral superior colliculus related to the initiation of pursuit and saccadic eye movements. *J. Neurosci.* 23, 4333–4344.

Heesy, C. P. (2009). Seeing in stereo: the ecology and evolution of primate binocular vision and stereopsis. *Evol. Anthropol.* 18, 21–35.

Hilbig, H., Bidmon, H. J., Ettrich, P., and Müller, A. (2000). Projection neurons in the superficial layers of the superior colliculus in the rat: a topographic and quantitative morphometric analysis. *Neuroscience* 96, 109–119.

Ignashchenkova, A., Dicke, P. W., Haarmeier, T., and Their, P. (2004). Neuron-specific contribution of the superior colliculus to overt and covert shifts of attention. *Nat. Neurosci.* 7, 56–64.

Krauzlis, R. J., Lovejoy, L. P., and Zénon, A. (2013). Superior colliculus and visual spatial attention. *Annu. Rev. Neurosci.* 36, 165–182.

Kustov, A. A., and Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature* 384, 74–77.

Landry, R., and Bryson, S. E. (2004). Impaired disengagement of attention in young children with autism. *J. Child Psychol. Psychiatry* 45, 1115–1122.

Kobayashi, T., Tran, A. H., Nishijo, H., Ono, T., and Matsumoto, G. (2003). Contribution of hippocampal place cell activity to learning and formation of goal-directed navigation in rats. *Neuroscience* 117, 1025–1035.

McHaffie, J. G., Stanford, T. R., Stein, B. E., Coizet, V., and Redgrave, P. (2005). Subcortical loops through the basal ganglia. *Trends Neurosci.* 28, 401–407.

McPeek, R. M., and Keller, E. L. (2004). Deficits in saccade target selection after inactivation of superior colliculus. *Nat. Neurosci.* 7, 757–763.

Müller, J. R., Philiastides, M. G., and Newsome, W. T. (2005). Microstimulation of the superior colliculus focuses attention without moving the eyes. *Proc. Natl. Acad. Sci. U.S.A.* 102, 524–529.

Munoz, D. P., and Istvan, P. J. (1998). Lateral inhibitory interactions in the intermediate layers of the monkey superior colliculus. *J. Neurophysiol.* 79, 1193–1209.

List of Changes

Replaced: ~~skill~~ replaced with: skills, on page 1, line 10.

Deleted: [date/time, etc.] precipitation, on page 2, line 15.

Added: all, on page 2, line 15.

Replaced: ~~hence~~ replaced with: therefore, on page 2, line 18.

Added: abbreviated as, on page 2, line 22.

Deleted: ~~In other words, probabilistic biases, which can be reduced by statistical post-processing methods, decrease over time.~~, on page 3, line 48.

Added: predictive, on page 8, line 86.