Redefining the Ensemble Spread-Skill Relationship from a Probabilistic Perspective

Eric P. Grimit and Clifford F. Mass
Department of Atmospheric Sciences
University of Washington, Seattle, WA USA

1. The Ensemble Spread-Skill Relationship

- Based on the premise that ensemble spread should be related to forecast error.
- Often characterized by the correlation between the standard deviation of the ensemble forecasts and the absolute error of the ensemble mean forecast – the spread-error correlation.
- Actual ensemble spread-error correlations have been disappointing, often less than 0.6, especially for short-range ensemble forecasts (SREF) of near-surface weather parameters.
- Larger spread-error correlations (0.6-0.7) have been observed with the aid of spatial-averaging, forecast bias correction, and considering only extreme spread cases (Grimit and Mass 2002; Stensrud and Yussouf 2003).
- However, correlations of spatial averages can be misleading (ecological correlation) and the practical use of spatial-mean error forecasts is suspect.
- Simple, statistical arguments (Houtekamer 1993) show that ensemble spread-error correlation is a function of the day-to-day ensemble spread variability (β) and thus has an upper limit near 0.8.

$$\rho(\sigma, |E|) = \sqrt{\frac{2}{\pi} \cdot \frac{1 - \exp(-\beta^2)}{1 - \frac{2}{\pi} \cdot \exp(-\beta^2)}}. \qquad \lim_{\beta \to \infty} \rho(\sigma, |E|) = \sqrt{\frac{2}{\pi}}.$$

- A couple studies show that categorical measures of forecast spread (e.g., statistical entropy or mode population) are better than continuous measures (e.g., variance) at discriminating forecast successes from failures (Toth et al. 2001; Ziehmann 2001).
  - NCEP developed the Relative Measure of Predictability (RMOP) based on these results
  - Can this be verified with a toy model of the spread-skill relationship?

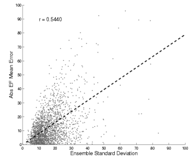
2. A Probabilistic Perspective

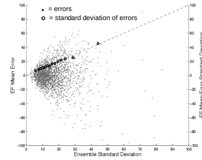
- What is the best performance we can expect given a perfect ensemble of finite size?
- Using a modified version of the Houtekamer (1993) toy statistical model (see inset) to generate large numbers of idealized ensemble forecasts, it is observed that spread-error data distributions are not linear.
- While small spread guarantees a small error, large spread does not guarantee a large error. There is simply a higher probability of observing a large error.
- A probabilistic approach that uses the full forecast error distribution is warranted.
- Looking at the ensemble-mean forecast error distribution in groups organized by ensemble spread, it is easy to notice that the spread in the errors is coincident with the ensemble spread.
- Therefore, a perfect spread-skill relationship can be redefined as a higher-order statement of statistical consistency, where ensemble variance equals ensemble-mean error variance over all classes of ensemble spread, not just in aggregate.

Modified Statistical Model of Spread-Skill

1. Draw today's "forecast uncertainty" from a log-normal distribution (Houtekamer 1993 model).
   $\ln(\sigma) \sim N(\ln(\sigma_f), \beta^2)$
2. Explicitly simulate ensemble forecasts by drawing M values from the ideal forecast PDF.
   $X_m \sim N(Z, \sigma^2)$; $m = 1,2,\ldots,M$
3. Let the verifying observation be drawn from the same distribution to ensure statistical consistency.
   $Y \sim N(Z, \sigma^2)$

- Repeat above for 10^4 realizations (cases)
- Assumed:
  - Gaussian statistics
- Varied:
  - temporal spread variability (β)
  - ensemble size (M)
  - spread and error metrics (continuous and categorical)





3. Idealized Forecast Error Prediction

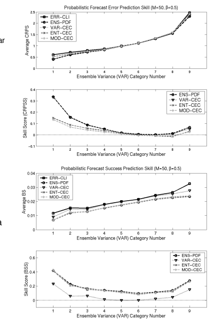
- The spread-skill relationship can be quantified in terms of how much ensemble spread information improves probabilistic error forecasts.
- Historical ensemble-mean forecast error distributions, conditioned by ensemble spread category, form the probabilistic forecast error predictions:

  Spread-based conditional error climatology (CEC)

  [Note: think of this approach as "dressing" the ensemble-mean forecast with its historical error distribution from cases with similar ensemble forecast spread.]
- Skill is measured relative to the overall, unconditional error climatology (ERR-CLI; level of no spread-skill relationship) and the direct ensemble probability density function (ENS-PDF; perfect probabilistic error forecast).
- All error forecasts are smoothed by fitting Gaussian parameters (mean and variance).
- A cross-validation approach is used to form the error climatologies from independent training data for each historical test case.
- The continuous ranked probability score (CRPS) and its associated skill score (CRPSS) are used to evaluate the probabilistic error forecasts from the perspective of a user with a continuous cost function.
- To evaluate skill from the perspective of an end-user with a categorical cost function:
  - Ensemble forecasts and verifications are divided among 10 climatologically equally likely categories.
  - Success is defined as the event that the verification and the ensemble mean fall into identical categories.
  - The Brier score (BS) and its associated skill score (BSS) are used to evaluate the probabilistic success forecasts.

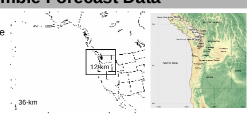
Table 5.1. CRPSS over the entire 10^4-case sample for the ENS-PDF, VAR-CEC, ENT-CEC, and MOD-CEC methods of probabilistic forecast error prediction as a function of the number of forecast spread categories.

No. Spread Bins	ENS-PDF	VAR-CEC	ENT-CEC	MOD-CEC
3	0.055	0.041	0.017	0.014
5	0.058	0.051	0.026	0.018
7	0.060	0.054	0.027	0.020
9	0.060	0.055	0.027	0.021

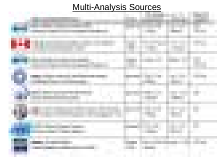


4. UW Short-Range Ensemble Forecast Data

- The UW Mesoscale Ensemble (UWME)
  - A multianalysis-based short-range ensemble forecast (SREF) system for the Pacific Northwest using MM5 at 36- and 12-km
  - Eight members, each with unique initial conditions and lateral boundary conditions taken from the large-scale analyses and forecasts of major operational weather centers worldwide
  - A parallel, eight-member ensemble that includes physics diversity (UWME+)
- Evaluation Period:
  - 361 cases (31 OCT 2002 – 31 MAR 2004) !!!
    - 271 cool season (Oct.-Mar.) cases
    - 90 warm season (Apr.-Sep.) cases
    - All forecasts initialized at 0000 UTC
    - 48h duration, stored every 3h
- Parameters of Focus [assumed distribution]:
  - 2-m Temperature (T_2) [Gaussian]
  - 10-m Wind Speed ($WSPD_{10}$) [Gaussian after √ - transform]
  - 10-m Wind Direction ($WDIR_{10}$) [von Mises]
- Verification Data:
  - NCEP Rapid Update Cycle 20-km Analysis (RUC20)
    - 12-km SREF fit to RUC20 domain over common area
  - Observations
    - 12-km SREF bi-linearly interpolated to observation sites



Multi-Analysis Sources



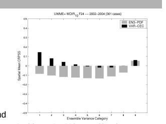
MM5 Physics Configurations for UWME and UWME+

IC	ID#	PBL	PBL: Soil	PBL: vertical diffusion	Cloud Microphysics	Cumulus: 36-km Domain	Cumulus: 12-km Domain	Cumulus: shlw. cumls	Radiation	SST Perturbation	Land Use Table
UWME		MRF	5-Layer	Y	Simple Ice	Kain-Fritsch	Kain-Fritsch	N	cloud	standard	standard
UWME+											
gfs	plus01	MRF	LSM	Y	Simple Ice	Kain-Fritsch	Kain-Fritsch	Y	RRTM	SST_pert01	LANDUSE.plus1
cmcg	plus02	MRF	5-Layer	Y	Reisner II	Grell	Grell	N	cloud	SST_pert02	LANDUSE.plus2
eta	plus03	Eta	5-Layer	N	Goddard	Betts-Miller	Grell	Y	RRTM	SST_pert03	LANDUSE.plus3
gasp	plus04	MRF	LSM	Y	Shultz	Betts-Miller	Kain-Fritsch	N	RRTM	SST_pert04	LANDUSE.plus4
jma	plus05	Eta	LSM	N	Reisner II	Kain-Fritsch	Kain-Fritsch	Y	cloud	SST_pert05	LANDUSE.plus5
ngps	plus06	Blackadar	5-Layer	Y	Shultz	Grell	Grell	N	RRTM	SST_pert06	LANDUSE.plus6
tcwb	plus07	Blackadar	5-Layer	Y	Goddard	Betts-Miller	Grell	Y	cloud	SST_pert07	LANDUSE.plus7
ukmo	plus08	Eta	LSM	N	Reisner I	Kain-Fritsch	Kain-Fritsch	N	cloud	SST_pert08	LANDUSE.plus8

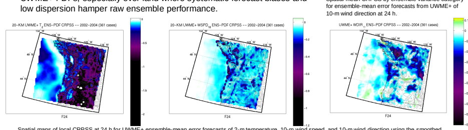
1) Albedo
2) Roughness Length
3) Moisture Availability

5. Real Forecast Error Prediction

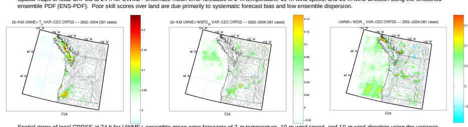
- Using the 361-case archive of real mesoscale SREF data from UWME and UWME+ forecasts, the ensemble-spread skill relationship can be assessed for individual grid points.
- On average, VAR-CEC forecasts make a 0-5% improvement over the error climatology (ERR-CLI) depending on the parameter and forecast lead time.
- As quantified by the skill of variance-based conditional error climatology (VAR-CEC) forecasts, strong ensemble spread-skill relationships for near-surface wind and temperature are found to be highly localized.
- In contrast to the idealized situation, ERR-CLI forecasts beat the UWME and UWME+ PDFs, especially over land where systematic forecast biases and low dispersion hamper raw ensemble performance.



Spatial-mean CRPSS by ensemble variance category for ensemble-mean error forecasts from UWME+ of 10-m wind direction at 24 h.



Spatial maps of local CRPSS at 24 h for UWME+ ensemble-mean error forecasts of 2-m temperature, 10-m wind speed, and 10-m wind direction using the smoothed ensemble PDF (ENS-PDF). Poor skill scores over land are due primarily to systematic forecast bias and low ensemble dispersion.



Spatial maps of local CRPSS at 24 h for UWME+ ensemble-mean error forecasts of 2-m temperature, 10-m wind speed, and 10-m wind direction using the variance-based conditional error climatology (VAR-CEC) method. VAR-CEC skill scores measure the ensemble spread-skill relationship.

6. Summary and Conclusions

- Does the spread of mesoscale short-range ensemble forecasts contain reliable information from which the expected errors in near-surface weather forecasts can be estimated *a priori*?
  - Previous attempts to answer this involved the use of linear correlations, sometimes with spatial-averaging.
  - After noting that correlations of spatial-averages can be misleading and that predictions of spatial-mean forecast errors have little utility for most end users, this study focused on the prediction of local forecast errors.
  - Using a toy, statistical model, it is shown that ensemble spread-error correlation is an inadequate description of the spread-skill relationship and does not appropriately account for the uncertainties in forecast error prediction.
  - A perfect spread-skill relationship can be redefined as a higher-order statement of statistical consistency, where ensemble variance equals the ensemble-mean error variance over all classes of ensemble spread.
- The predictive skill of spread-based conditional error climatologies (CECs) is proposed as a robust measure of the redefined spread-skill relationship.
  - For idealized ensembles where Gaussian error statistics apply, spread-based CEC skill is nearly equal to the skill of the smoothed ensemble PDF (ENS-PDF).
  - End users should choose a spread metric consistent with their own cost function (either continuous or categorical) to form the most appropriate CECs.
- Obtaining skillful, local forecast error predictions of near-surface temperature and wind was difficult.
  - Even when physics diversity is used (UWME+), temporal spread variability is low.
  - Major problems with the MM5 physics and surface boundary parameters tend to produce errors that swamp any spread-skill relationship over most of the domain.
  - In localized areas (and during specific periods) where the forcing by terrain is strong enough to overwhelm the physics errors, a significant spread-skill relationship is observed.
  - For example, many areas with a strong spread-skill relationship for wind direction forecasts also tend to be locations that often experience channeled flows.
  - Forecast bias and low dispersion render direct ensemble PDFs less skillful than error climatology, especially over land.
  - A simple forecast bias correction applied to each ensemble member individually using a 14-day sliding window training period improves the ensemble forecast probabilities, but degrades the ensemble spread-skill relationship by reducing the temporal spread variability (not shown).

18-21 October 2004 **Workshop on Ensemble Methods** **Exeter, United Kingdom**

Supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745 and by a consortium of Federal and local agencies in the United States including the National Weather Service.