Current Issues and Challenges in Ensemble Forecasting

Carolyn Reynolds, Chiashi Muroi, and Tom Hamill

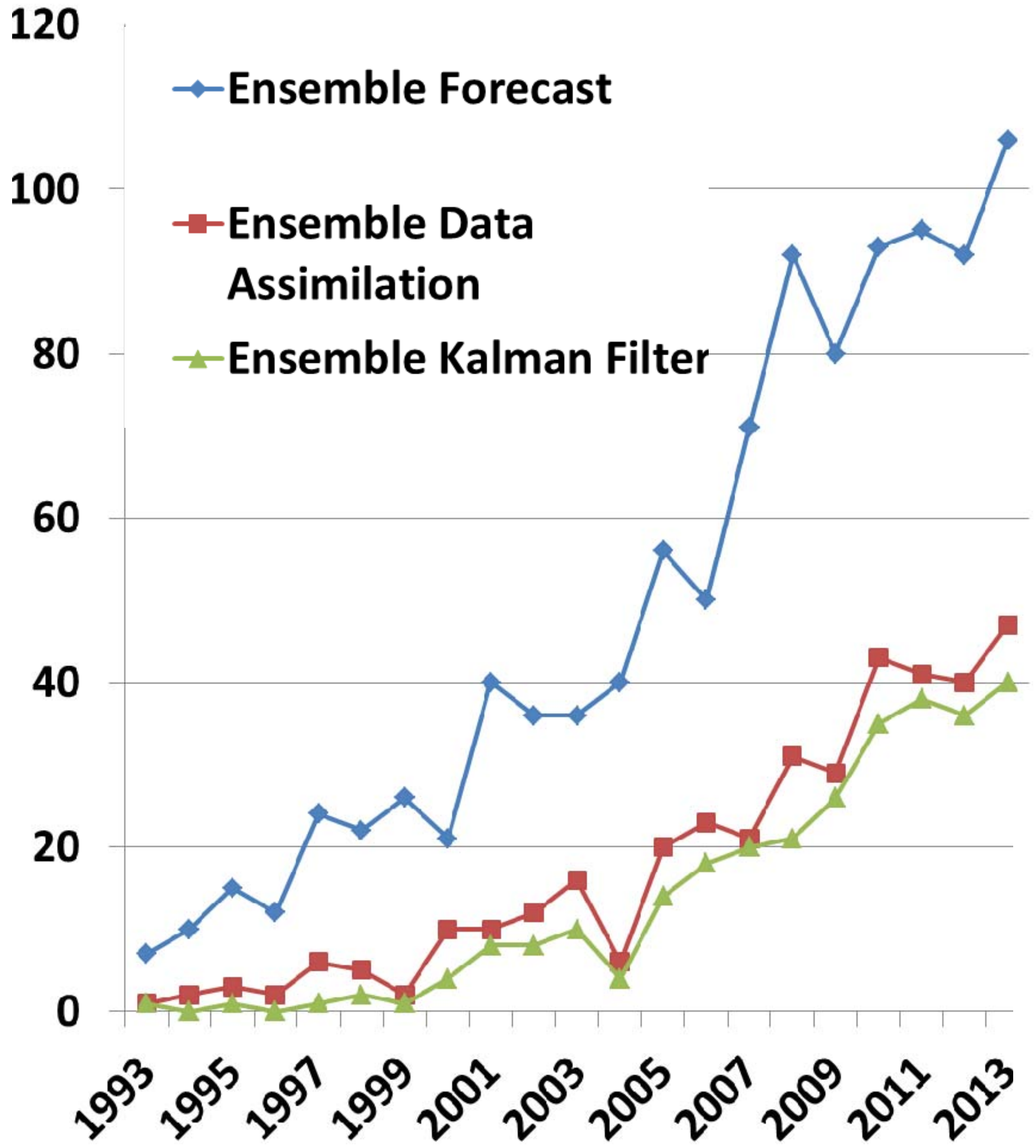
29th WGNE
Melbourne, Australia, 10-13 March 2014

- Recap of WGNE28
- Recent trends in ensemble-related research
- Accounting for model uncertainty
- Reforecasting and post-processing/calibration
- Multi-model ensemble issues and questions
- Coupling to other components (ocean, land surface)

Ensemble Summary from WGNE28 Report
Hamill and Muroi

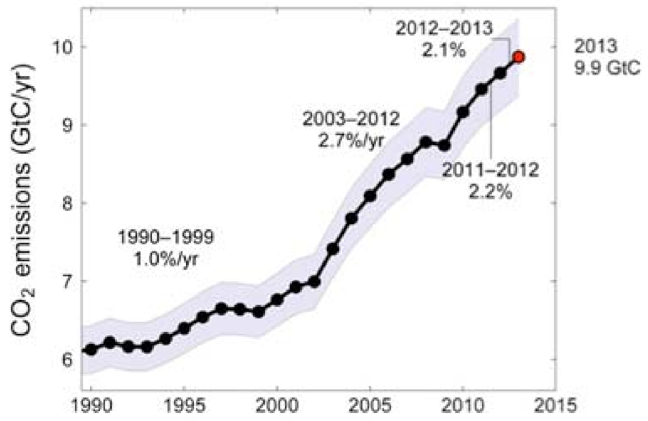
- **Centers moving towards initial perturbations representative of analysis errors**
  This continues, especially with spread of Ensemble Kalman Filter DA Techniques
- **Operational and Research centers developing methods for improved representation of uncertainty from model imperfections and inappropriate use of deterministic parameterizations** **Fast-growing interest in accounting for model uncertainty (PHY-EPS workshop Madrid 2013)**
- **Review of NOAA's GEFS reforecasts: statistical post-processing reduces forecast bias and improves reliability** **Other examples include Meteo-France 21-yr PEARP reforecasts, but reforecasting still not as widespread as it could/should be. Post-processing techniques have proliferated in the last several years.**
- **Increased development of cloud-permitting regional ensembles: How well is this community organized?** **Many examples of high-resolution ensembles exist, meetings such as PHY-EPS bring community together.**

*Number of Article/Year with These Words in the Abstract**



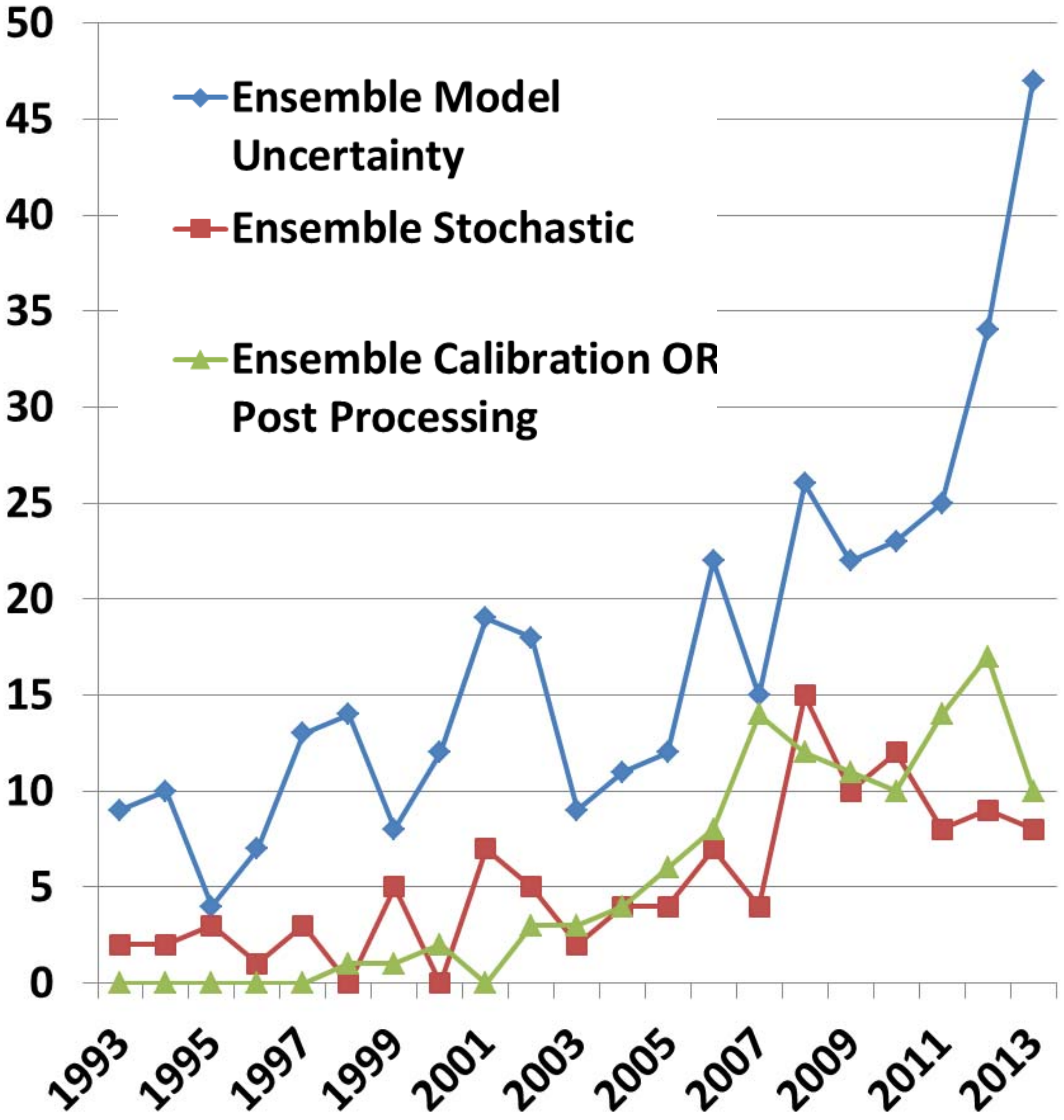
Research in ensemble forecasting and ensemble data assimilation has been climbing steadily.

Increasing even faster than C02 emissions!



*AMS journals only

*Number of Article/Year with These Words in the Abstract**

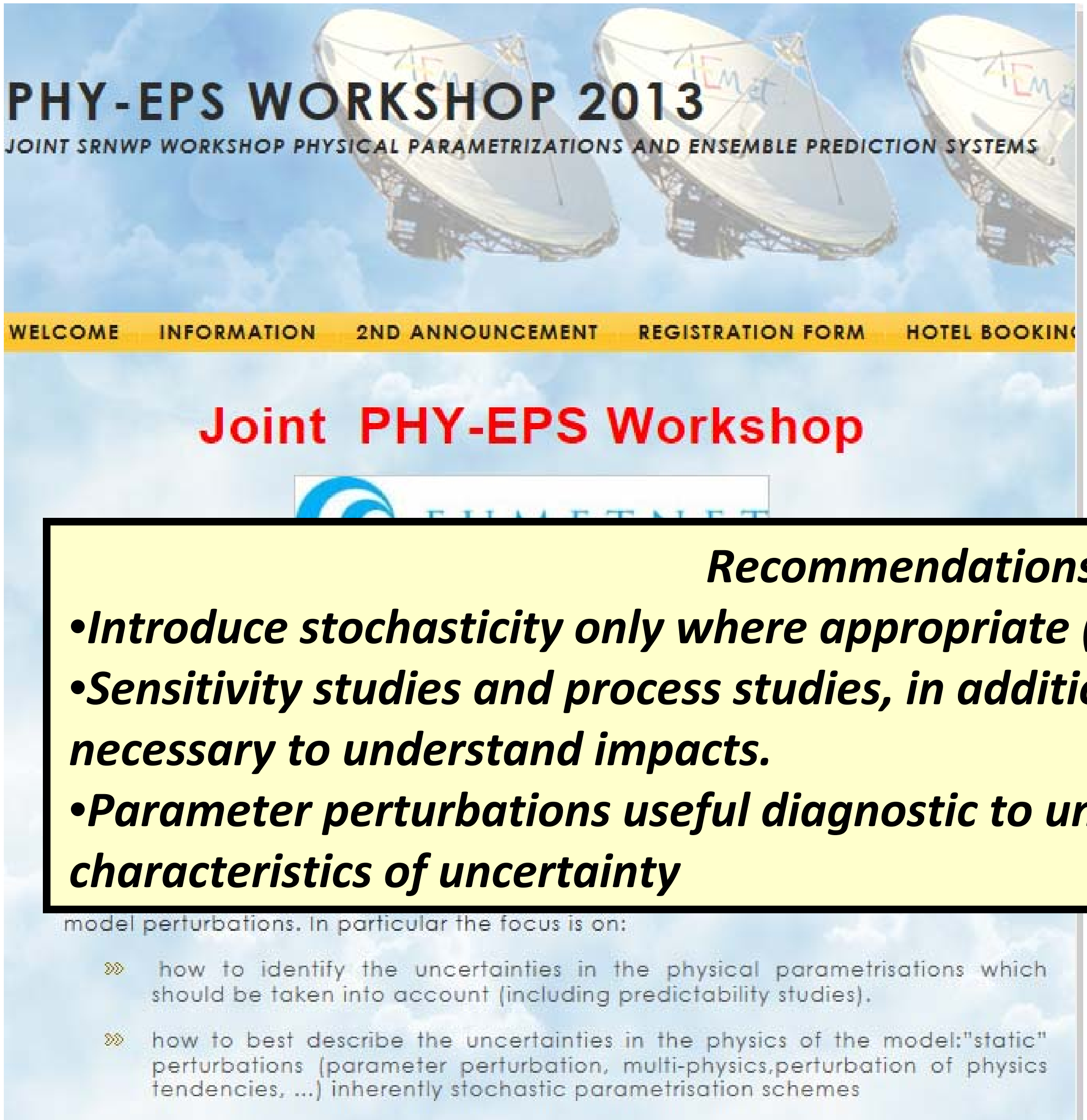


Research in Model Uncertainty grows rapidly in the last three years.

Interest in calibration and post-processing also substantially larger than in the early 2000s.

***AMS journals only**

Growing Interest in Accounting for Model Uncertainty



PHY-EPS WORKSHOP 2013

JOINT SRNWP WORKSHOP PHYSICAL PARAMETRIZATIONS AND ENSEMBLE PREDICTION SYSTEMS

Joint PHY-EPS Workshop

model perturbations. In particular the focus is on:

- how to identify the uncertainties in the physical parametrisations which should be taken into account (including predictability studies).
- how to best describe the uncertainties in the physics of the model:"static" perturbations (parameter perturbation, multi-physics,perturbation of physics tendencies, ...) inherently stochastic parametrisation schemes

Schemes: SKEB, SPPT, multi-physics, parameter variations, humidity perturbations, stochastic microphysics, stochastic convection

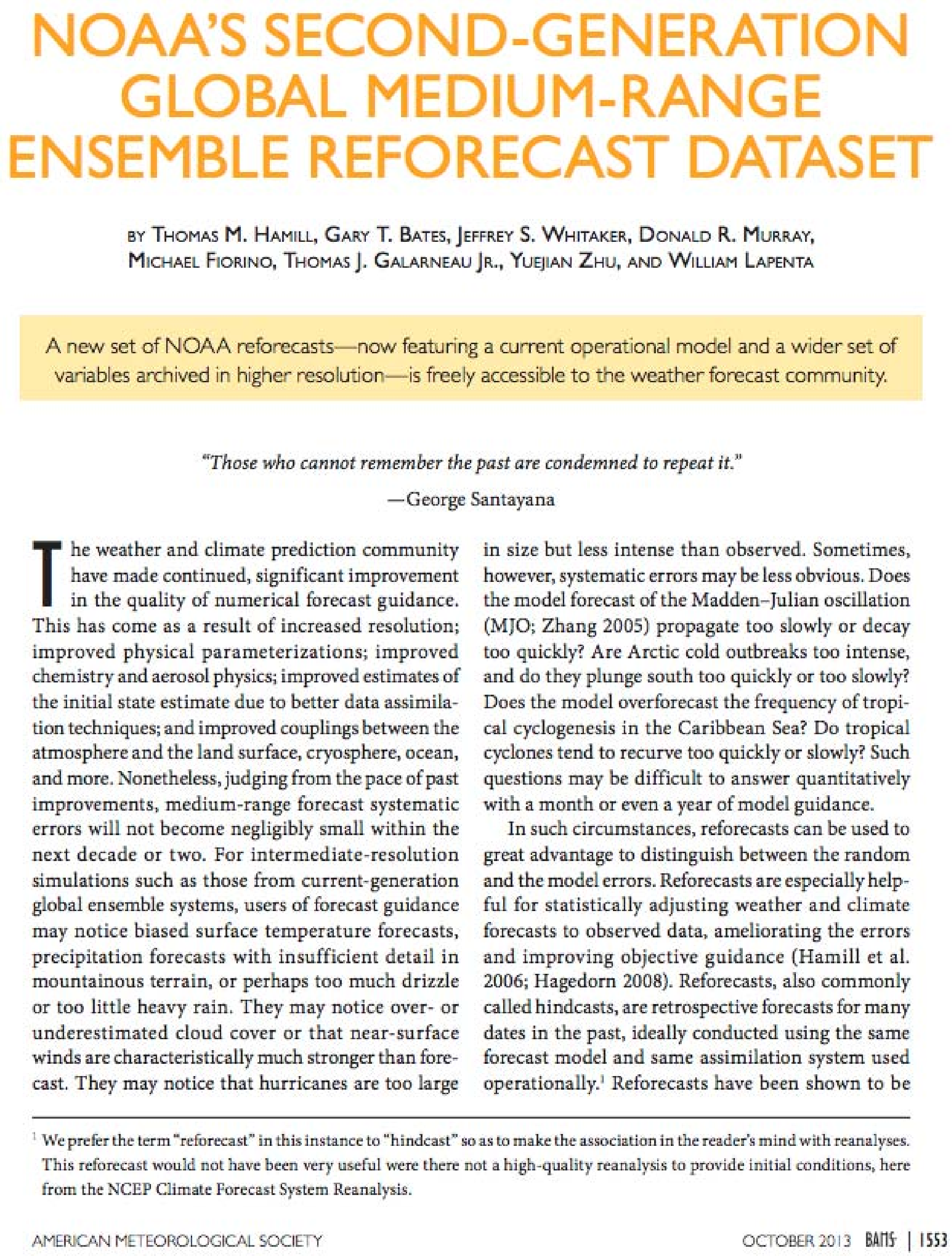
Mesoscale Models: WRF, MOGREPS, ALADIN/HARMONIE, AEROME, COSMO

Recommendations:

•Introduce stochasticity only where appropriate (maintain physical meaning).

•Sensitivity studies and process studies, in addition to predictability studies, are necessary to understand impacts.

•Parameter perturbations useful diagnostic to understand spatio-temporal characteristics of uncertainty



NOAA'S SECOND-GENERATION GLOBAL MEDIUM-RANGE ENSEMBLE REFORECAST DATASET

BY THOMAS M. HAMILL, GARY T. BATES, JEFFREY S. WHITAKER, DONALD R. MURRAY, MICHAEL FIORINO, THOMAS J. GALARNEAU JR., YUEJIAN ZHU, AND WILLIAM LAPENTA

A new set of NOAA reforecasts—now featuring a current operational model and a wider set of variables archived in higher resolution—is freely accessible to the weather forecast community.

"Those who cannot remember the past are condemned to repeat it."

—George Santayana

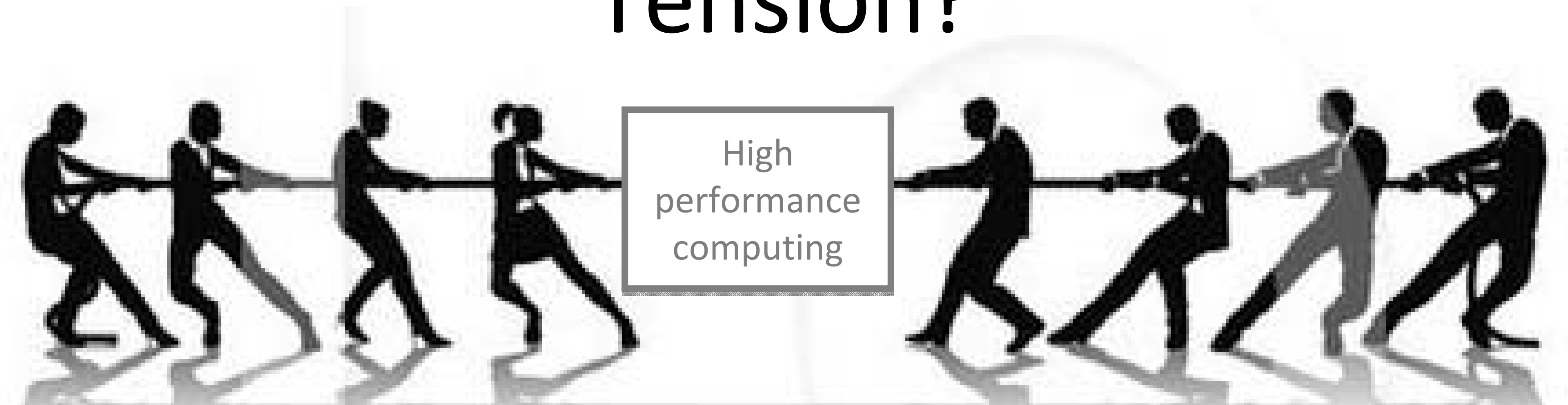
The weather and climate prediction community have made continued, significant improvement in the quality of numerical forecast guidance. This has come as a result of increased resolution; improved physical parameterizations; improved chemistry and aerosol physics; improved estimates of the initial state estimate due to better data assimilation techniques; and improved couplings between the atmosphere and the land surface, cryosphere, ocean, and more. Nonetheless, judging from the pace of past improvements, medium-range forecast systematic errors will not become negligibly small within the next decade or two. For intermediate-resolution simulations such as those from current-generation global ensemble systems, users of forecast guidance may notice biased surface temperature forecasts, precipitation forecasts with insufficient detail in mountainous terrain, or perhaps too much drizzle or too little heavy rain. They may notice over- or underestimated cloud cover or that near-surface winds are characteristically much stronger than forecast. They may notice that hurricanes are too large in size but less intense than observed. Sometimes, however, systematic errors may be less obvious. Does the model forecast of the Madden–Julian oscillation (MJO; Zhang 2005) propagate too slowly or decay too quickly? Are Arctic cold outbreaks too intense, and do they plunge south too quickly or too slowly? Does the model overforecast the frequency of tropical cyclogenesis in the Caribbean Sea? Do tropical cyclones tend to recurve too quickly or slowly? Such questions may be difficult to answer quantitatively with a month or even a year of model guidance.

In such circumstances, reforecasts can be used to great advantage to distinguish between the random and the model errors. Reforecasts are especially helpful for statistically adjusting weather and climate forecasts to observed data, ameliorating the errors and improving objective guidance (Hamill et al. 2006; Hagedorn 2008). Reforecasts, also commonly called hindcasts, are retrospective forecasts for many dates in the past, ideally conducted using the same forecast model and same assimilation system used operationally.[1] Reforecasts have been shown to be

[1] We prefer the term "reforecast" in this instance to "hindcast" so as to make the association in the reader's mind with reanalyses. This reforecast would not have been very useful were there not a high-quality reanalysis to provide initial conditions, here from the NCEP Climate Forecast System Reanalysis.

Reforecasting benefits well documented (NOAA GFS, ECMWF, ARPEGE, COSM-LEPS*), however, issues remain concerning broader adoption.

Tension?



Higher-resolution models, more models, run more frequently

Improved assimilation methods

Improved physics

Frequent model updates

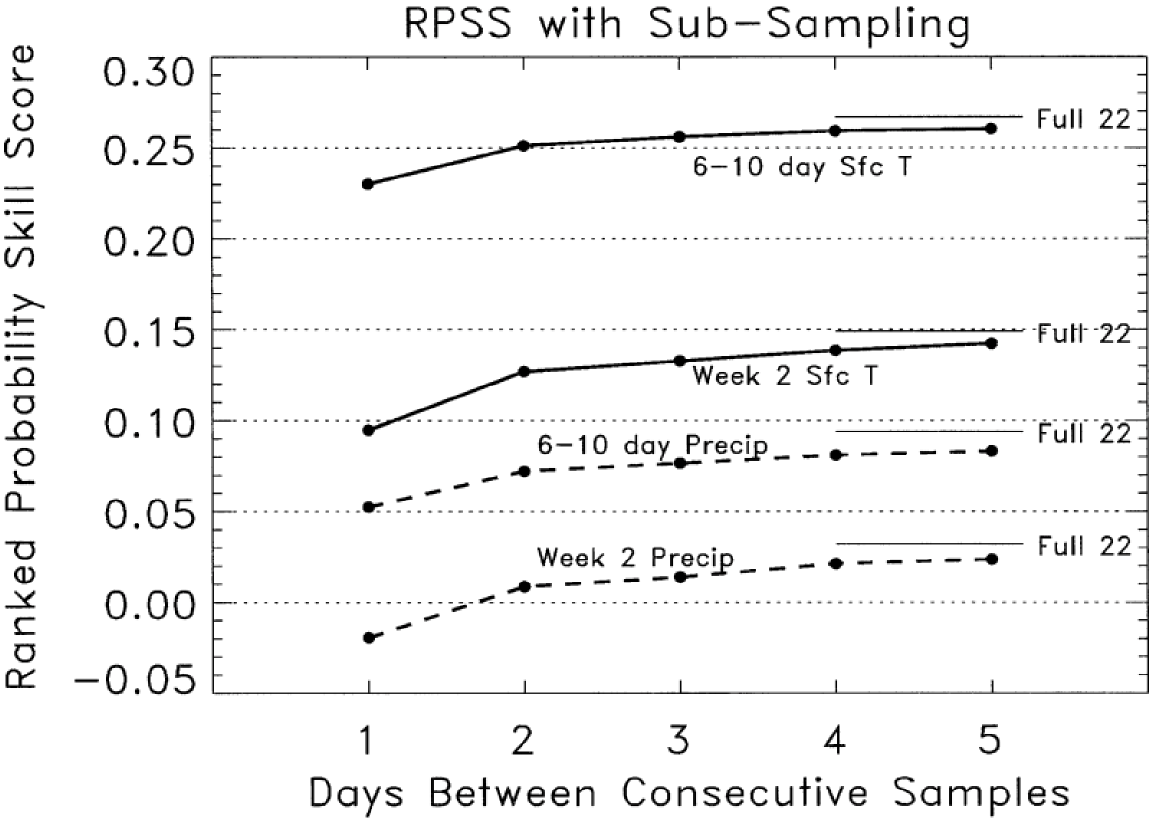
More ensemble members

Retrospective forecasts

Reanalyses to initialize retrospective forecasts

More stable models

Are there ways to decrease the number of reforecasts needed?



Yes, we think.

Here, four years of reforecast data are computed, with up to 5 days between consecutive samples. Spacing out the reforecasts provides almost as much post-processed skill as training with 22 years of every-day reforecast data for these applications.

Ref: Hamill et al., MWR, 2004

Reforecast Challenges

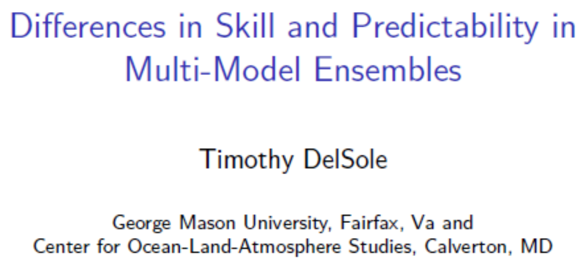
- Reforecasts require past initial conditions with accuracy like that of real-time analyses. Hence, regular reanalyses needed.
  - Also, ensembles of initial conditions generated in the same manner as real-time ensemble.
- Ensemble systems such as the SREF that use different models, different physics may have larger reforecast requirements than systems with “exchangeable” members. We may need a reforecast for each member, with its unique biases, or need to rethink the SREF configuration.

Number of multi-model ensembles are growing

Mesoscale: TIGGE-LAM, NOAA SREF, AEMET-SREPS, SESAR, CAPS, HFIP
Global 1-2 weeks: NAEFS, NUOPC, TIGGE, HIWPP
Subseasonal to seasonal: NMME, DEMETER,

Why do multi-model ensemble often outperform single model ensembles? Is the improvement in skill due to larger ensemble size or to combining signals?

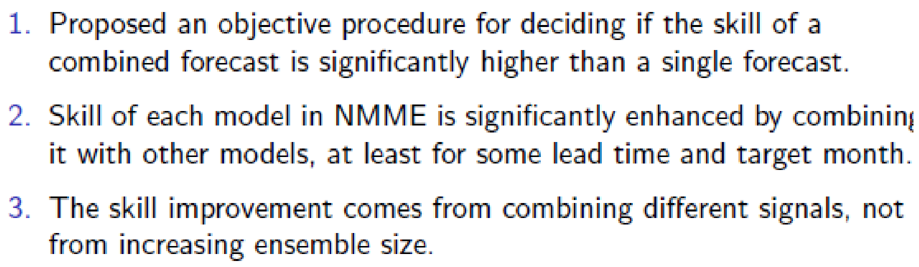
International Conference on S2S prediction, 10-13 Feb 2014



Differences in Skill and Predictability in Multi-Model Ensembles

Timothy DelSole

George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD



1. Proposed an objective procedure for deciding if the skill of a combined forecast is significantly higher than a single forecast.
2. Skill of each model in NMME is significantly enhanced by combining it with other models, at least for some lead time and target month.
3. The skill improvement comes from combining different signals, not from increasing ensemble size.

- ***How does one combine multi-model forecasts of unequal skill? Equal weights competitive with more complex schemes (DelSole et al. 2012, Sansom et al. 2013, …)***

- ***Tradeoffs between independence from multi-models vs. focusing resources on one system.***

- ***Issues of latency, data transfer reliability, etc.***

SESAR EU project: test multi-ensembles (for aviation)

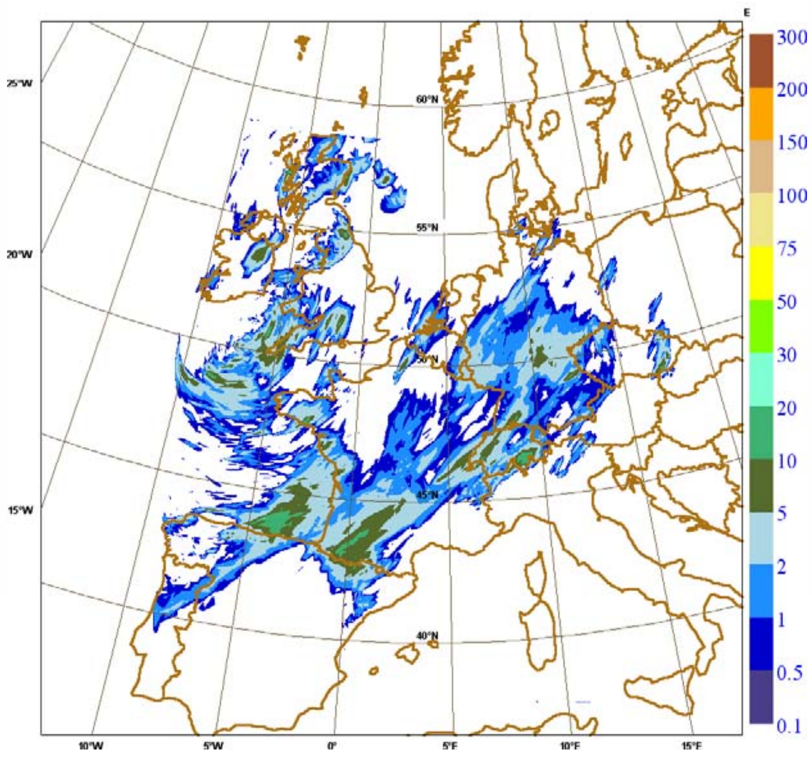
Arome+UKV+COSMODE = 12+12+20 members = 44 members

all members have equal weight (except for PDF smoothing near each model's domain edge) & similar resolution ~ 2.5km

Objective scores: multi-ensemble usually better than each ensemble in the overlap zones.

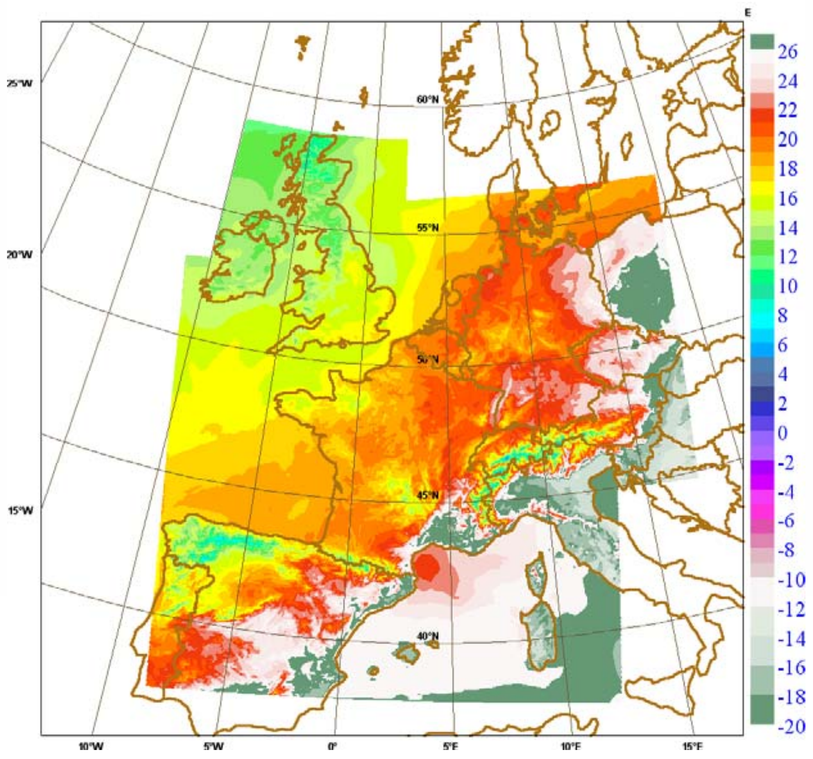
3-model PDF precip

Mean prec9h(mm) 2012080500+09



3-model PDF of T2m

Mean T2m(C) 2012080500+09



More centers are testing coupling to (or incorporating uncertainty from) other components of the earth system

Examples include:

- **ECMWF land surface perturbations**
- **Met Office GloSea ocean coupling**
- **Meteo France AROME EPS surface pertrubations**
- **DWD COSMO-DE-EPS: Perturbed soil moisture**
- **CMC: surface and near surface model error representation.**
- **NRL Coupled COAMPS ET, SST perturbations for NAVGEM**
- **NCEP: by 2018, one/two way coupling with ocean model and perturbed land surface**

Potential Issues: Peter Houtekamer notes that as system become coupled (land/ocean/atmosphere/ice), they become increasingly more complex. Ensuring that coupled systems out-perform un-coupled ones will require close collaboration amongst the groups working on the different components.

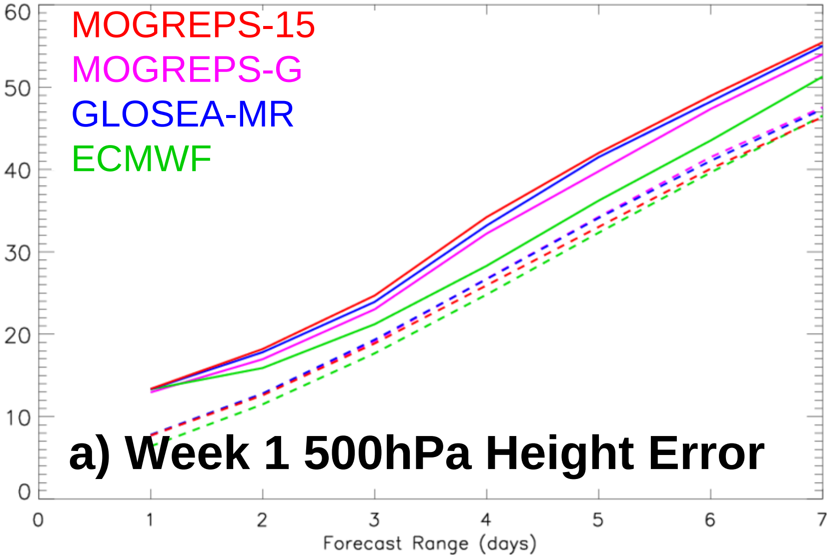
Extra slides

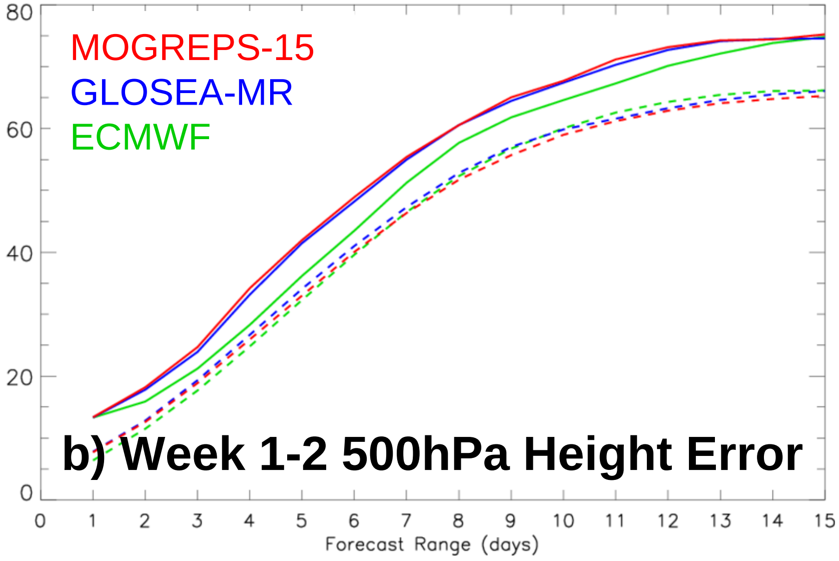
GloSea-Medium Range (MR) Project

Met Office

- Assess sensitivity of short-/medium range ensemble forecast skill to:
  - Resolution: N216 (60km) MOGREPS-15 vs N400 (~32km) MOGREPS-G **-> positive impact in week 1 shown in a)**
  - Ocean coupling: Atmosphere only MOGREPS-15 vs coupled O-A GLOSEA-MR **-> neutral impact in week 1-2 shown in a)-b).**

- Decisions: Do not implement GloSea-MR, retire MOGREPS-15, use ECMWF/multi-model for weeks2-4, extend MOGREPS-G to 7 days.

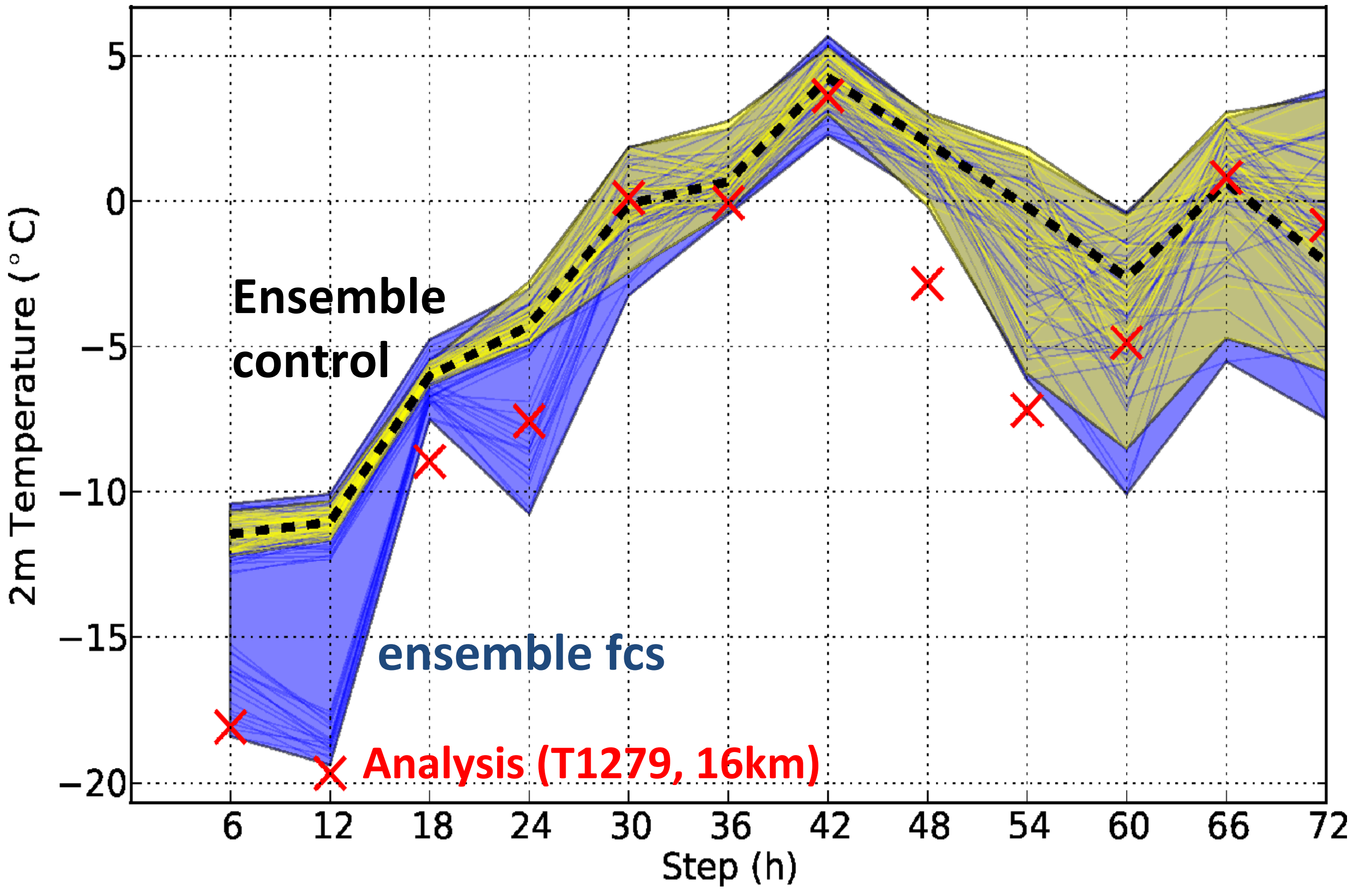

a) Week 1 500hPa Height Error


b) Week 1-2 500hPa Height Error

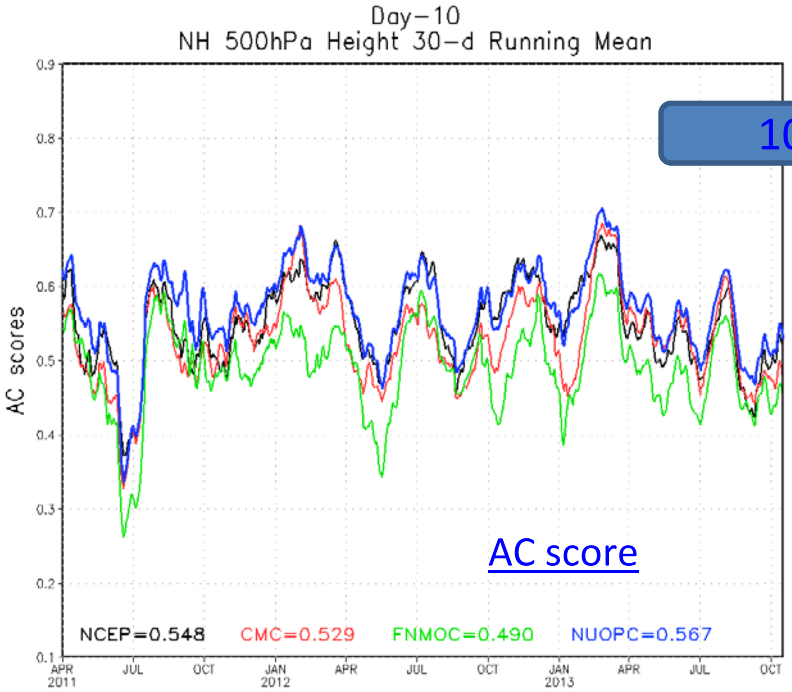
Cy40r1: EDA-based land-surface pert. in ENS ICs

Initial perturbations have been improved by stronger coupling of the ensemble of analyses and forecasts.

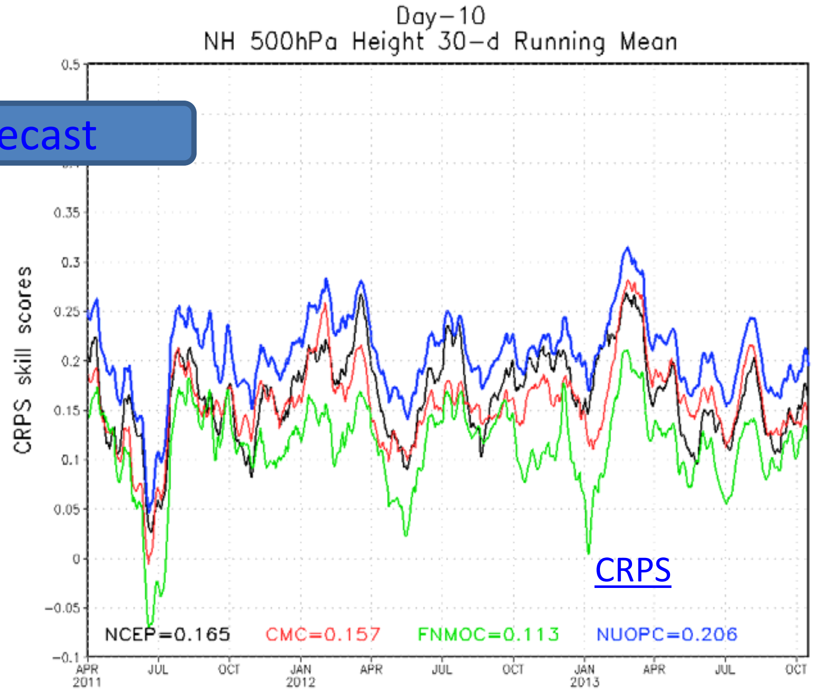
The perturbations to the land surface ICs increase the ENS spread of 2mT. The difference is largest in some situations, e.g. if there is uncertainty associated with snow cover which can lead to large differences in 2m temperature between the ensemble members.



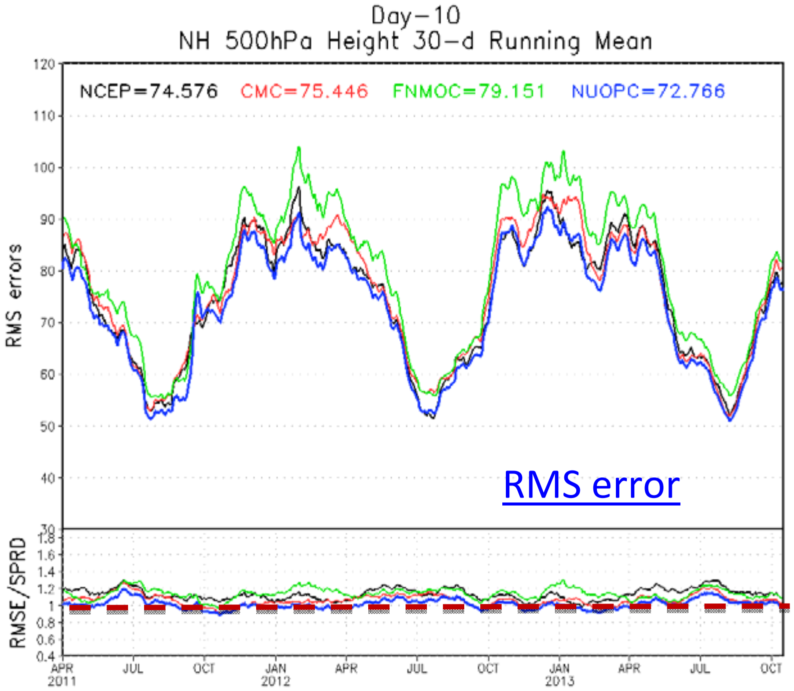
10-day forecast



AC score



CRPS



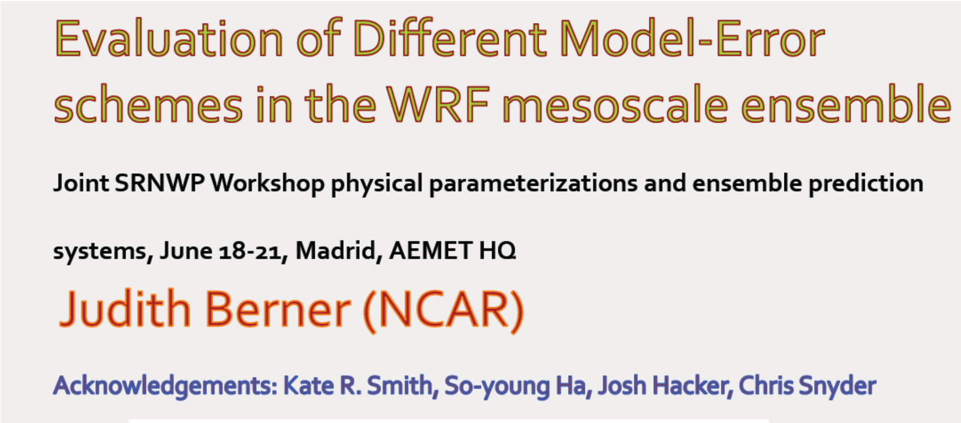
RMS error

Northern Hemisphere 500hPa height:

30-day running mean scores of day-10
CRPS skill score
RMS error and ratio of RMS error / spread
Anomaly correlation

All other regions could be seen from:
http://www.emc.ncep.noaa.gov/gmb/yluo/naefs/VRFY_STATS/T30_P500HGT

Different model uncertainty methods can be complementary (many examples) GET IMAGE



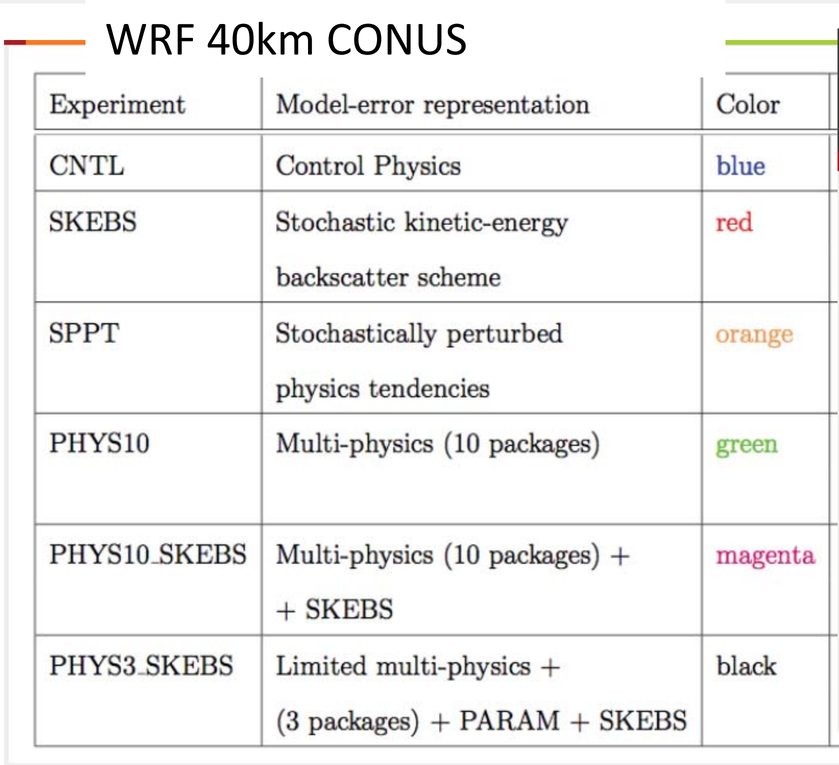
Evaluation of Different Model-Error schemes in the WRF mesoscale ensemble

Joint SRNWP Workshop physical parameterizations and ensemble prediction

systems, June 18-21, Madrid, AEMET HQ

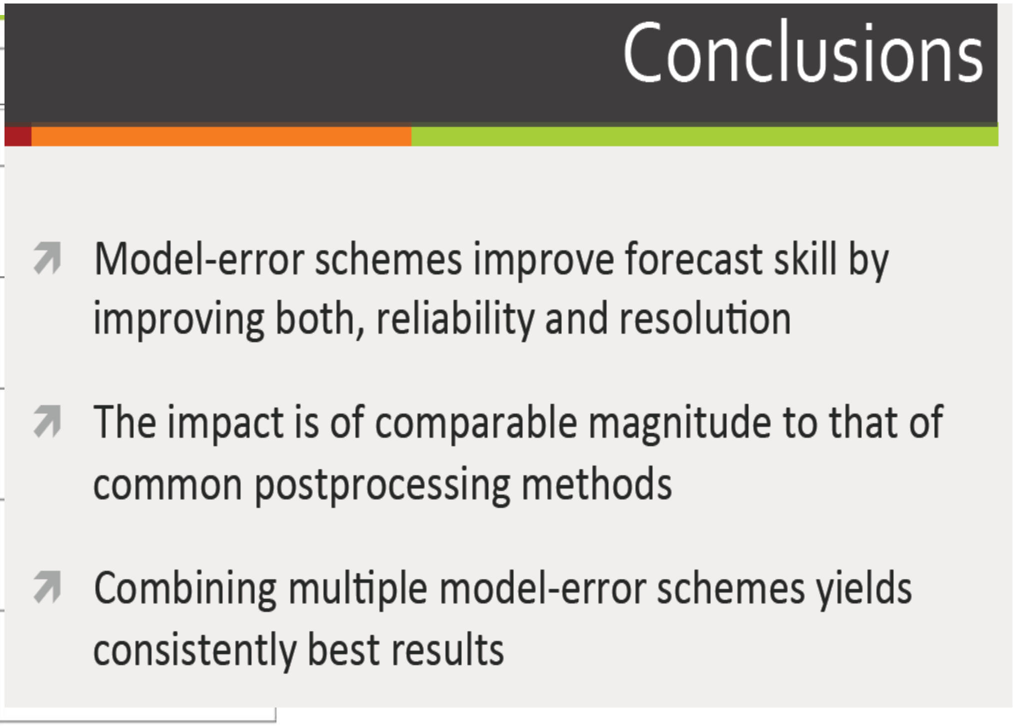
Judith Berner (NCAR)

Acknowledgements: Kate R. Smith, So-young Ha, Josh Hacker, Chris Snyder



WRF 40km CONUS

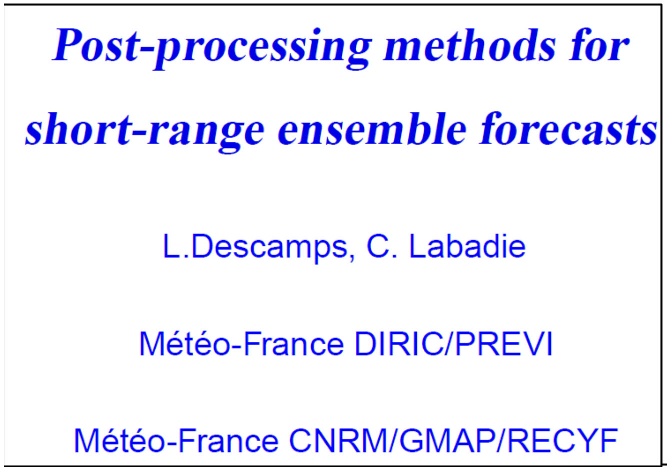
	Experiment	Model-error representation	Color
1	CNTL	Control Physics	blue
3	SKEBS	Stochastic kinetic-energy backscatter scheme	red
2	SPPT	Stochastically perturbed physics tendencies	orange
4	PHYS10	Multi-physics (10 packages)	green
6	PHYS10_SKEBS	Multi-physics (10 packages) + + SKEBS	magenta
5	PHYS3_SKEBS	Limited multi-physics + (3 packages) + PARAM + SKEBS	black



Conclusions

- Model-error schemes improve forecast skill by improving both, reliability and resolution
- The impact is of comparable magnitude to that of common postprocessing methods
- Combining multiple model-error schemes yields consistently best results

Advantages of post-processing with reforecasts well established (e.g., Hamill et al. 2013)

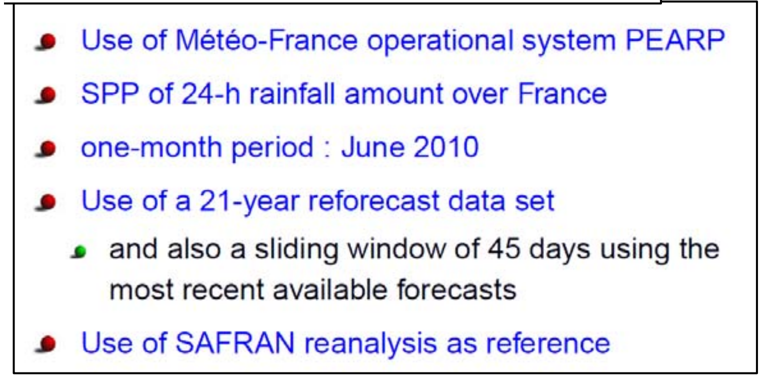


Post-processing methods for short-range ensemble forecasts

L.Descamps, C. Labadie

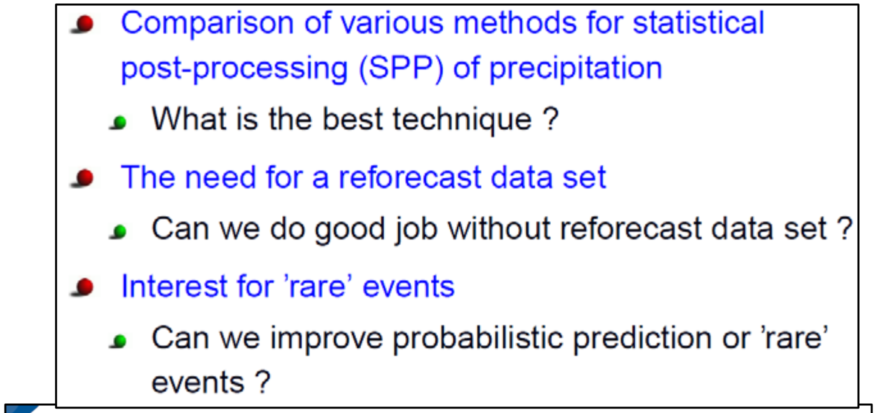
Météo-France DIRIC/PREVI

Météo-France CNRM/GMAP/RECYF

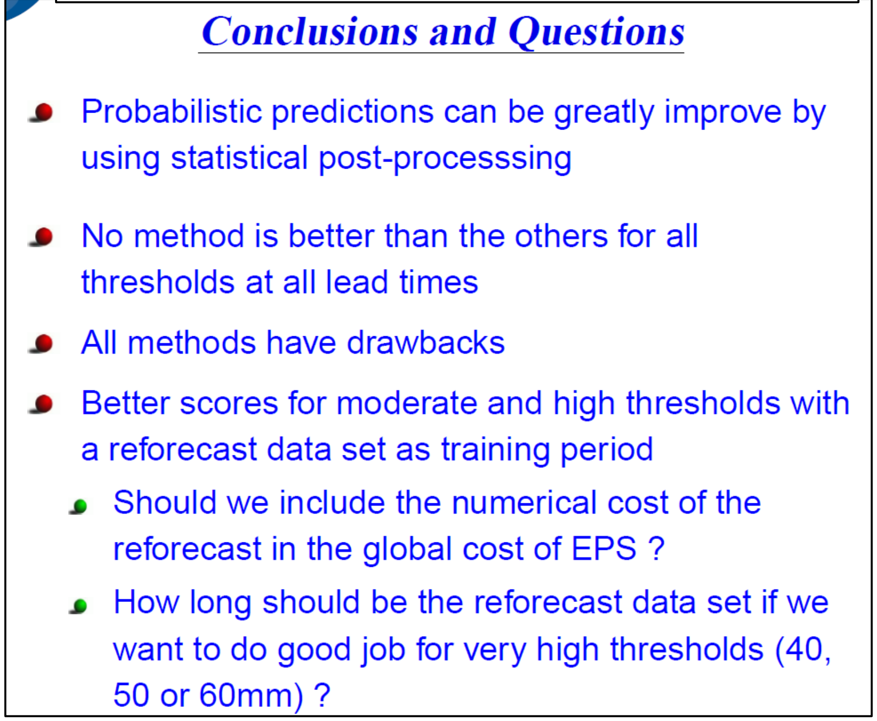


- Use of Météo-France operational system PEARP
- SPP of 24-h rainfall amount over France
- one-month period : June 2010
- Use of a 21-year reforecast data set
  - and also a sliding window of 45 days using the most recent available forecasts
- Use of SAFRAN reanalysis as reference

Techniques: Simple bias correction; CDF-based correction; Rank-Analog; Logistic Regression; BMA



- Comparison of various methods for statistical post-processing (SPP) of precipitation
  - What is the best technique ?
- The need for a reforecast data set
  - Can we do good job without reforecast data set ?
- Interest for 'rare' events
  - Can we improve probabilistic prediction or 'rare' events ?



Conclusions and Questions

- Probabilistic predictions can be greatly improve by using statistical post-processsing
- No method is better than the others for all thresholds at all lead times
- All methods have drawbacks
- Better scores for moderate and high thresholds with a reforecast data set as training period
  - Should we include the numerical cost of the reforecast in the global cost of EPS ?
  - How long should be the reforecast data set if we want to do good job for very high thresholds (40, 50 or 60mm) ?

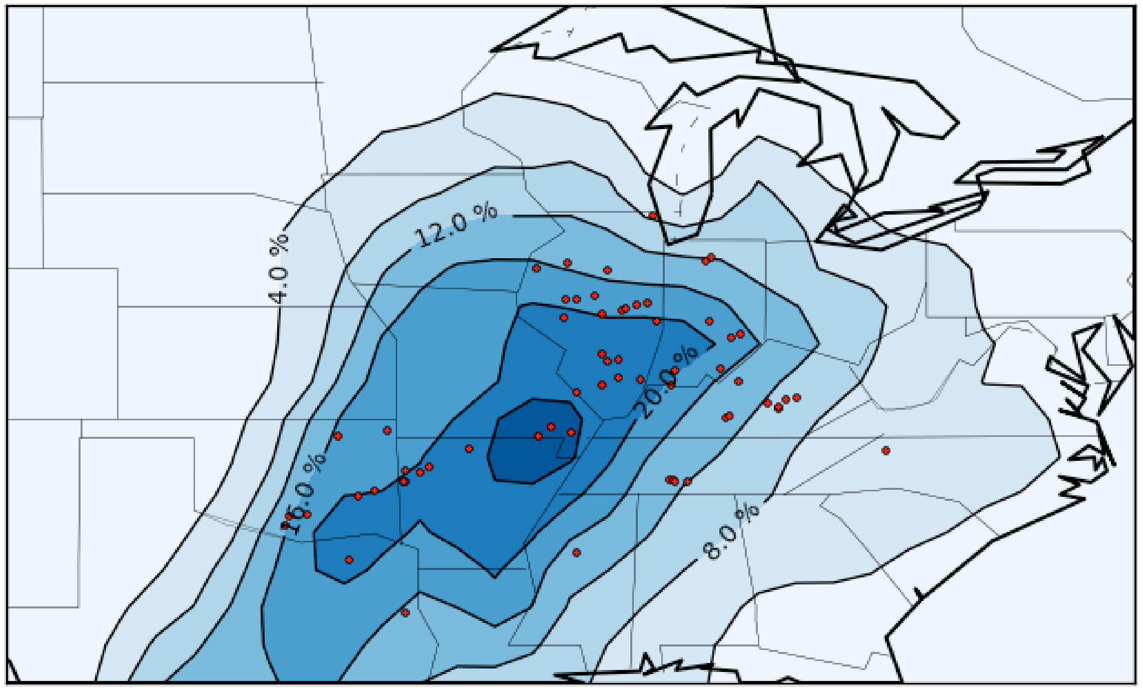
Post-processing techniques area of active research

Instituting Reforecasting at NCEP/EMC

Tom Hamill (ESRL)
Yuejian Zhu (EMC)
Tom Workoff (WPC)
Kathryn Gilbert (MDL)
Mike Charles (CPC)
Hank Herr (OHD)
Trevor Alcott (Western Region)

Exciting new products are also possible.

8.5 to 11.5 – day tornado forecast, 4/11/1996

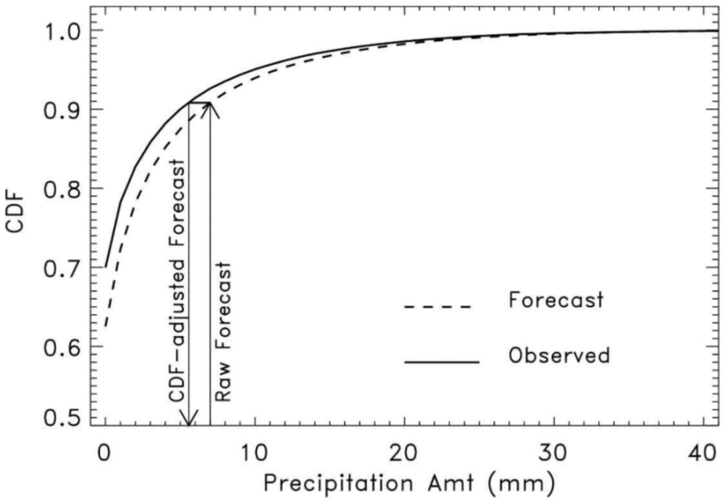


Francisco Alvarez at
St. Louis University,
is working with me
and others on using the
reforecasts to make
extended-range
predictions of
tornado probabilities.

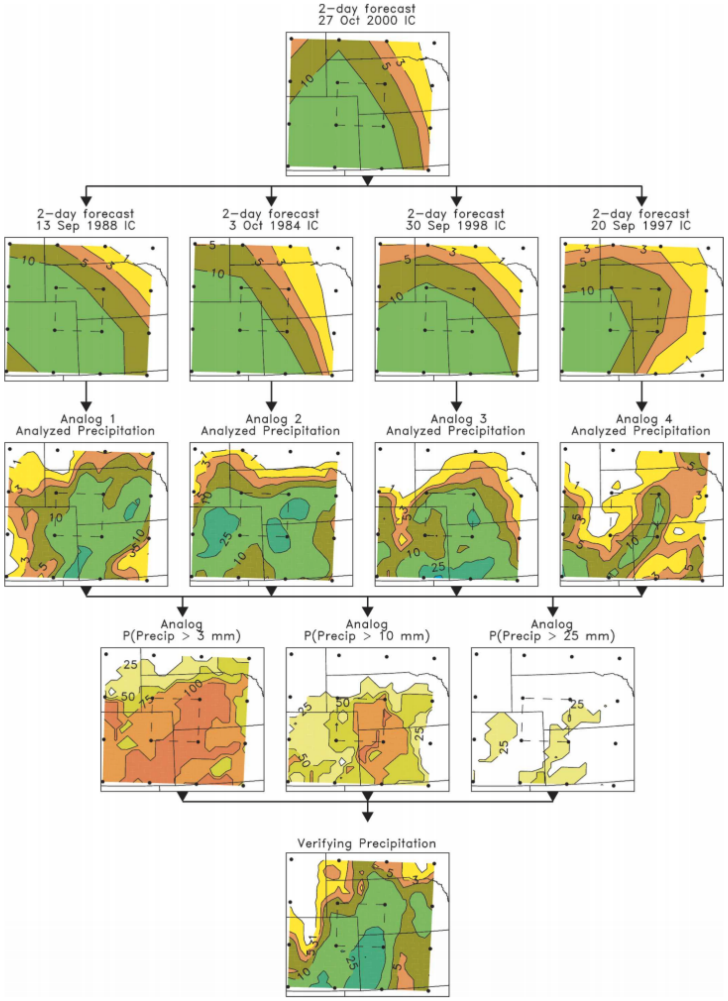
Ph.D. work,
in progress.

Many methods of post-processing.

CDF-based bias correction



Forecast analog



Ref: Hamill and Whitaker, MWR, 2006

Many post-processing methods, not all are equally skillful

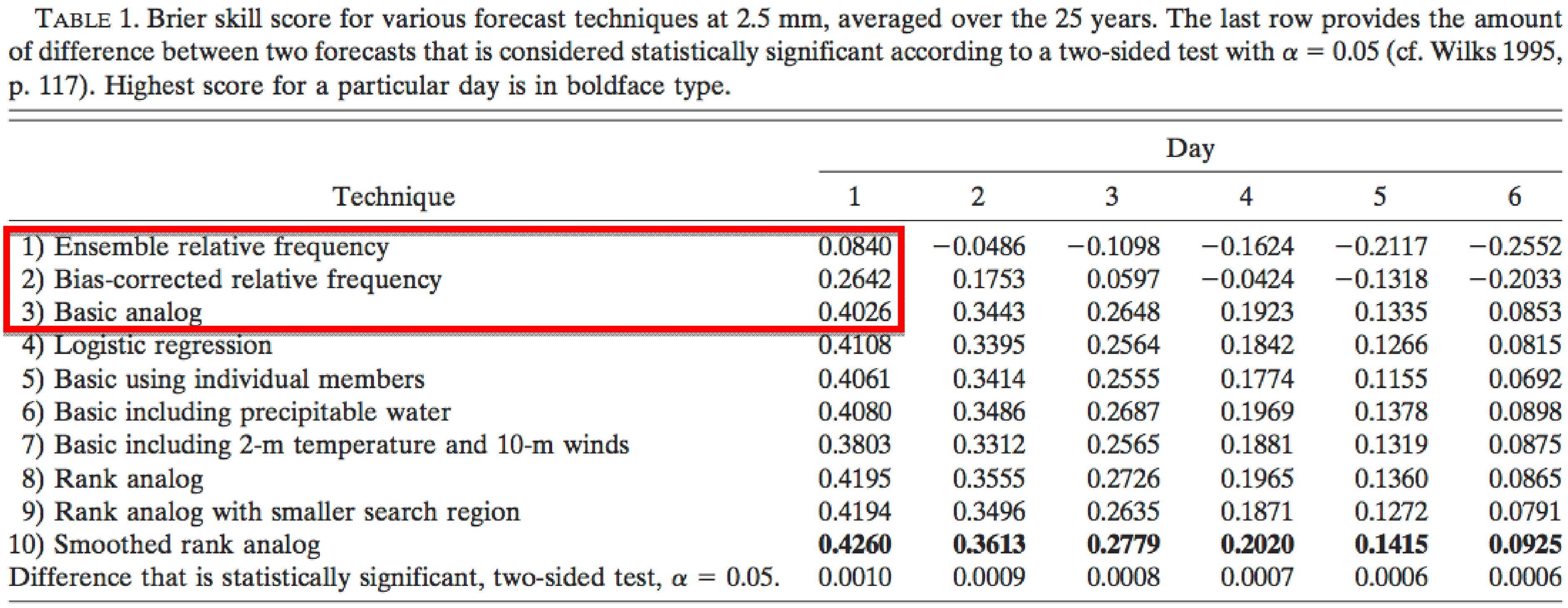


Table 1. Brier skill score for various forecast techniques at 2.5 mm, averaged over the 25 years. The last row provides the amount of difference between two forecasts that is considered statistically significant according to a two-sided test with $\alpha = 0.05$ (cf. Wilks 1995, p. 117). Highest score for a particular day is in boldface type.

	Day					
Technique	1	2	3	4	5	6
1) Ensemble relative frequency	0.0840	−0.0486	−0.1098	−0.1624	−0.2117	−0.2552
2) Bias-corrected relative frequency	0.2642	0.1753	0.0597	−0.0424	−0.1318	−0.2033
3) Basic analog	0.4026	0.3443	0.2648	0.1923	0.1335	0.0853
4) Logistic regression	0.4108	0.3395	0.2564	0.1842	0.1266	0.0815
5) Basic using individual members	0.4061	0.3414	0.2555	0.1774	0.1155	0.0692
6) Basic including precipitable water	0.4080	0.3486	0.2687	0.1969	0.1378	0.0898
7) Basic including 2-m temperature and 10-m winds	0.3803	0.3312	0.2565	0.1881	0.1319	0.0875
8) Rank analog	0.4195	0.3555	0.2726	0.1965	0.1360	0.0865
9) Rank analog with smaller search region	0.4194	0.3496	0.2635	0.1871	0.1272	0.0791
10) Smoothed rank analog	**0.4260**	**0.3613**	**0.2779**	**0.2020**	**0.1415**	**0.0925**
Difference that is statistically significant, two-sided test, $\alpha = 0.05$.	0.0010	0.0009	0.0008	0.0007	0.0006	0.0006

Post-processing skill often depends on training sample size.

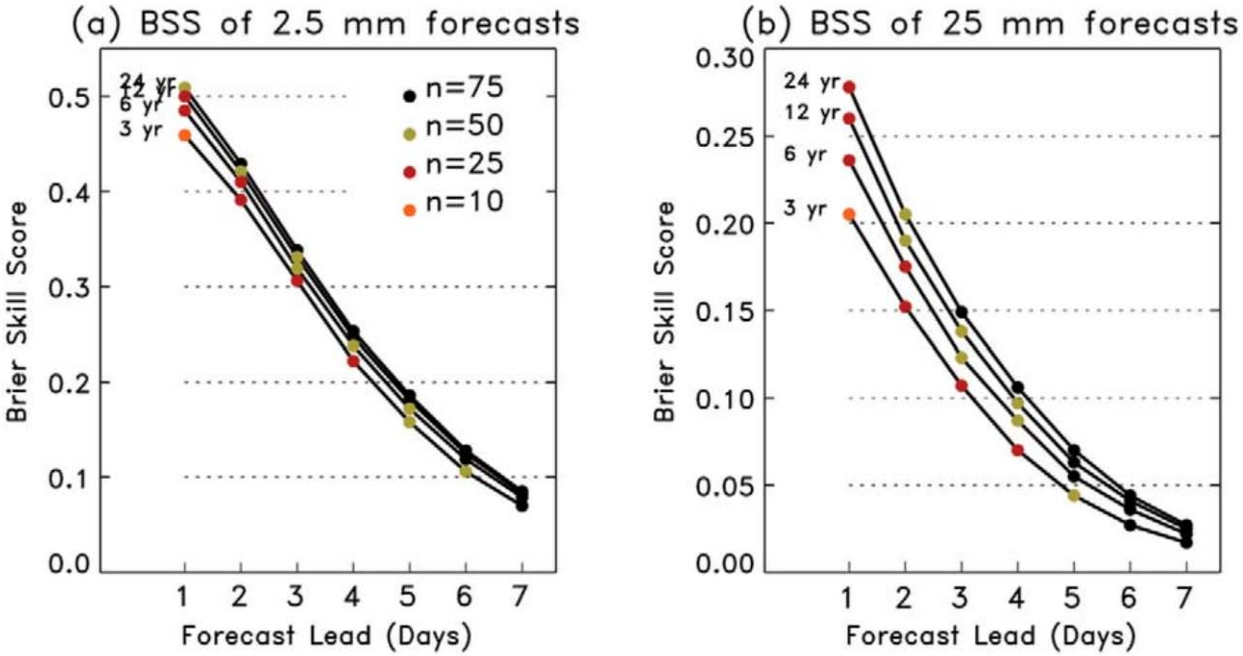


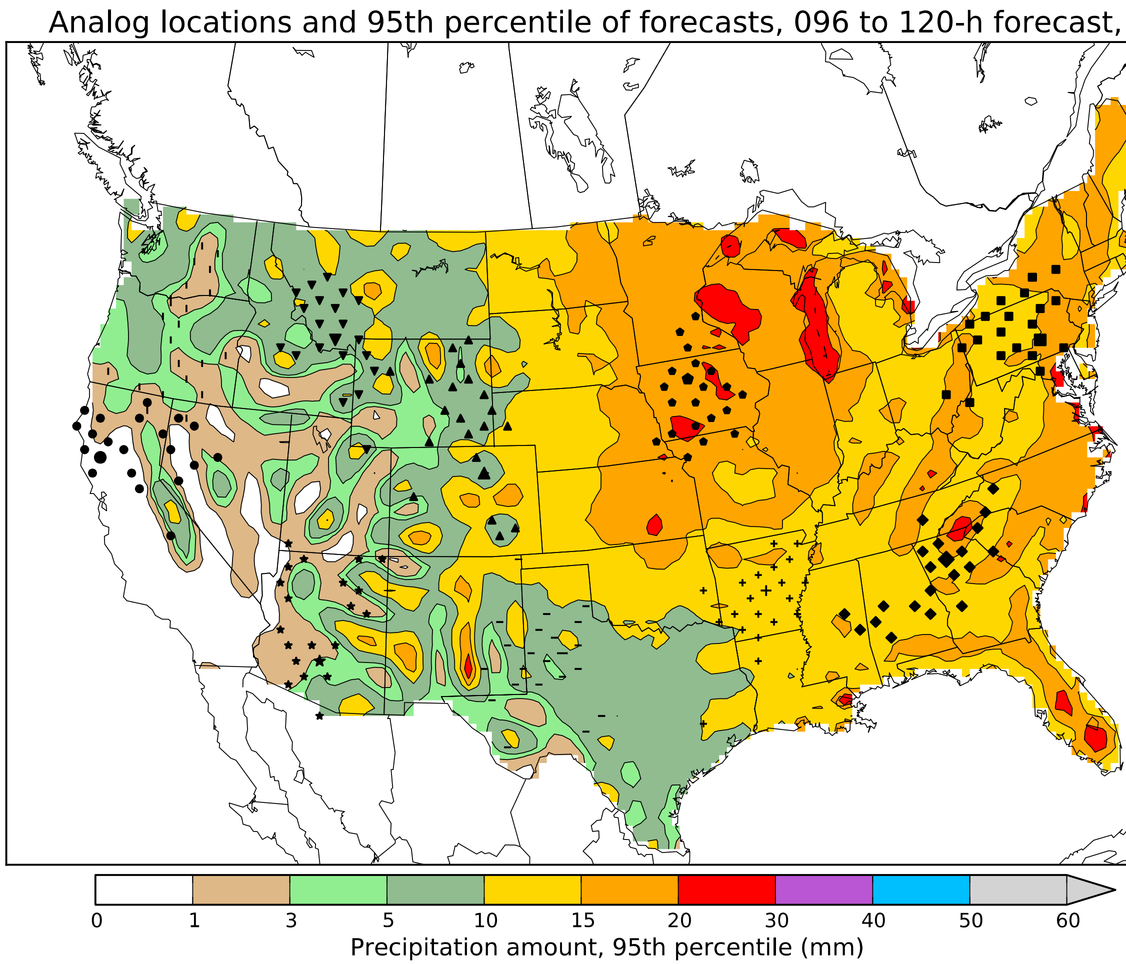
FIG. 7. Brier skill scores of the analog reforecast technique for various lengths of the training dataset. Probabilistic forecasts were calculated for ensembles of sizes 10, 25, 50, and 75; the skill of the ensemble size that was most skillful is the only one plotted. The color of the dot denotes the size of the most skillful ensemble.

There is more skill dependence on training sample size for the heavy precipitation (uncommon) than for light precipitation.

For many of the projects such as the “blender project” we are asked to calibrate variables such as precipitation that have this strong sample-size dependence.

Ref: Hamill et al. BAMS, 2006

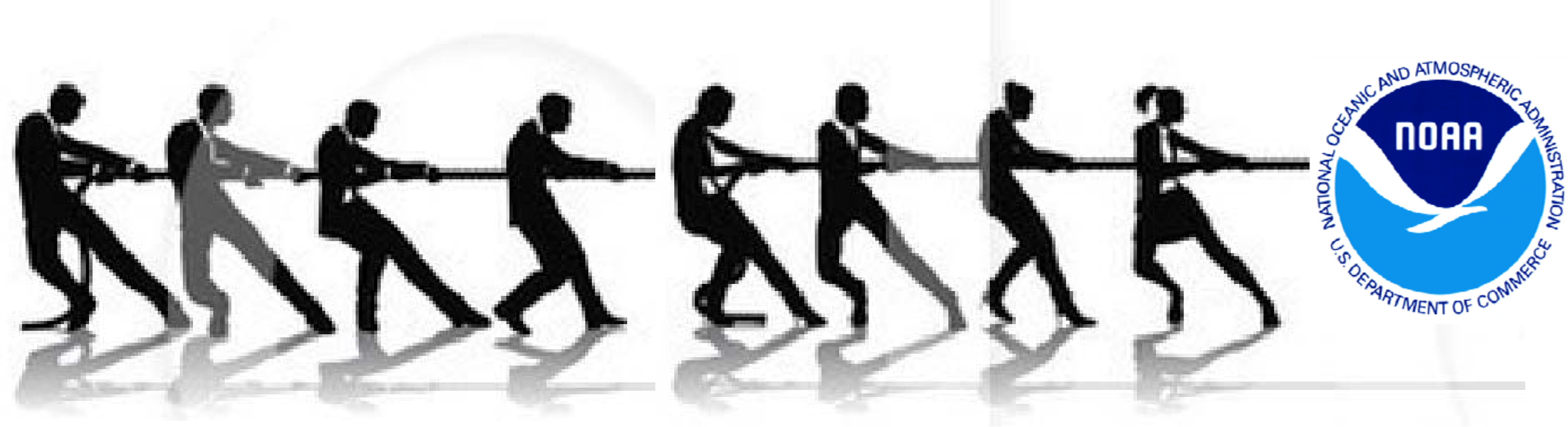
“Regionalization,” or training with data from supplemental locations can help (and hurt, too).


Analog locations and 95th percentile of forecasts, 096 to 120-h forecast, Jul
0
1
3
5
10
15
20
30
40
50
60
Precipitation amount, 95th percentile (mm)

Here, for a given grid point (big symbol) supplemental training data locations are identified that have similar forecast, observed climatologies. Approaches such as this can enlarge the training sample size, but sometimes forecast biases are very regionally specific, and this degrades the post-processing performance.

Ref: Hamill, yet unpublished work.

A mutual desired outcome?


NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA
U.S. DEPARTMENT OF COMMERCE

Rapidly improving models, assimilation methods, ensembles

An institutionalized, *light-footprint* reforecast capability to make the raw guidance even better