Reforecasts: An Important Dataset for Improving Weather Predictions

—Bob Glahn
Meteorological Development Laboratory,
Office of Science and Technology,
NOAA/National Weather Service,
Silver Spring, Maryland

Hamill et al. (2006, hereafter HWM) present a very interesting way to develop objective forecasts of surface weather variables. Basically, this consists of running some numerical model for a long sample period and developing statistical relationships between the weather elements and the archived output. The specific statistical method they present is an application of analogs in a model output statistics (MOS) framework; another method is presented in a similar paper by Hamill et al. (2004). They also suggest the forecasts, which they call "reforecasts," can be used to diagnose model bias and to study predictability.

The primary intent of the authors is"to stimulate a serious discussion about the value of reforecasts." I do not want to disappoint the authors and offer this as one "serious discussion." It is not meant to be critical, although I will present another point of view on some aspects.

The crucial question is whether a long record—here 25 yr—of a frozen model with appropriate initial conditions run at a lower resolution than the parent model, for economy, is of more use than a shorter record—maybe 5 yr—of an operational model that has possibly undergone some modest evolution over the period of record. I believe there are two main related considerations in trying to answer that question:

1) what is the ultimate purpose of the user of the reforecasts (e.g., diagnosing operational model behavior, studying predictability, making operational forecasts), and
2) can the stripped-down model furnish results competitive with an operational model?

Only by running a series of tests could the answer to 2) be determined, and even then it would only apply to the specific situation. It is possible that a long sample of low-resolution reforecasts would be more appropriate than a shorter sample of operational model forecasts for short-range forecasting than for longer range, or vice versa. For instance, the reforecast model will undoubtedly give good forecasts for the first day or two at the resolution at which it is run, but that resolution may not be optimal for short-range forecasts. For projections of a week or two, the resolution of the reforecasts would be quite sufficient for the predictable detail, but the model might not be accurate enough; for example, the wave speeds might suffer. HWM do not really address this question. They show results from a particular postprocessing technique for reforecasts, but only compare them to unpostprocessed National Centers for Environmental Prediction (NCEP) results (I do not consider just computing relative frequencies of the event from raw ensemble output postprocessing in the sense the term is usually used).

HWM have utilized two reanalysis datasets, the NCEP–National Center for Atmospheric Research (NCAR) global analysis (Kalnay et al. 1996) to initialize the NCEP Medium-Range Forecast (MRF) model run at T62 spectral resolution with 28 vertical levels (the same resolution at which the analyses were made), and the 32-km North American Regional Reanalysis (NARR; Mesinger et al. 2006) of precipitation data as the predictand dataset. While this paper is only one example of how a statistical system can be built from model data, if a "weather" forecast at projections under, say, 3 days is desired, the 32-km data of NARR cannot be considered of fine enough resolution to resolve "local variations of rainfall" (HWM, p. 38) or permit "the extraction of small scale detail" (HWM, p. 39) in today's world of gridded forecasts of 1.25- to 5-km resolution (Glahn and Ruth 2003).

DOI:10.1175/2008BAMS2725.1

We at the Meteorological Development Laboratory are finding that terrain features at 5 km are not of sufficient detail in rugged terrain to meet the needs of Weather Forecast Offices of the National Weather Service. Also, in making detailed statistical weather forecasts, it is not obvious that the predictand should be an analysis rather than the actual observations. However, such an analysis may be adequate and appropriate for longer-range predictions if finescale detail is not needed.

As with any statistical forecast system, several decisions have to be made in the course of development. For instance, HWM used a 91-day window in finding analogs, and the number of analogs to use in the final step of determining a mean or distribution has to be decided. After the system has been defined and forecasts made for the period of record, skill measures can be computed. According to HWM's Fig. 7, a 24-yr record of data is always better than a shorter one, the amount depending on the projection and the number of analogs. At 7 days there is not much difference, especially between a 12- and 24-yr record. Also, it appears that the best number of analyses varies with projection and number of years in the ensemble. In a way, it is perplexing that 75 analyses give higher skill at 7 days than a lesser number when only 3 yr of data are used (presumably $91 \times 3 = 273$ cases to chose from), but this possibly relates to the lack of forecastable detail at that projection. The more analogs chosen, the less some of them will resemble the actual forecast patterns, and that will tend to "spread" the precipitation (in this case) and make the forecast less specific. (I suppose in the caption "ensembles of sizes" is another use in the paper of "ensembles," which means "number of analogs.")

HWM mention (footnote 2, p. 36) "slight discontinuities" at the boundaries of the tiles (the local analog areas). We have found that techniques that use regions in which the same statistics have been applied may produce discontinuities that are unacceptable as a gridded product. This is a basic problem that must be addressed, now that many forecast products are displayed graphically (map form) at high resolution. There is no magic bullet; there is usually a trade-off between smoothing out wanted detail along with nonmeteorological discontinuities and noise. Hamill and Whitaker (2006) have now demonstrated one possibility that shows promise; we are investigating others.

HWM show results in their Fig. 5 and discuss the Brier skill score (BSS) without making it clear that the skill is relative to climatic relative frequencies. Equation (1) is shown, but it is not apparent how the calculation is made, and reference is made to an unpublished paper. Also perplexing is the statement concerning Fig. 8, "In general, the BSS tended to be smaller in drier regions, where the reference climatology is more skillful." How can climatology be more skillful, or skillful at all, if the same climatology is used as a reference? It is true that the raw Brier scores are smaller (better) in drier regions, as is the Brier score for climatology, and there is some tendency for the BSS to be smaller in direr regions, but not greatly so (Glahn and Jorgensen 1970). The definition of "climatology" is extremely important, and even in Eq. (1) given by Wilks (2006) the "uncertainty" is many times quite gross, being calculated over some long period. I assume that however skill was defined it was computed in the same way for "NCEP Opnl" and "New" (Fig. 5), so the comparison is appropriate. It is not surprising that most any reasonable postprocessing technique will improve on raw relative frequencies from an ensemble. Of more interest is whether a specific method improves on existing postprocessing methods.

It is interesting that the ensemble mean precipitation forecast pattern was used to find analogs rather than using each ensemble member individually in some way. HWM state this individual use produced less skillful forecasts (footnote, p. 36). This is a relatively minor use of the concept of ensembles of model forecasts. The real use of ensembles here is in using the reanalysis precipitation patterns.

On the subject of diagnosing operational model bias from reforecasts, I question whether a long record of a model run at lower resolution, and maybe with other differences, than an operational model will yield more useful results than a shorter record of the "real thing," even if the real thing had some evolutionary changes. Detail like that in HWM's Fig. 9 does need a long sample (bias computed by lead time and day with a 31-day window), but how would one know that a large portion of the bias was not just due to the decrease in model resolution?

The use of principal components (PCs), and to a lesser extent canonical correlation, to study teleconnections and predictability over large regions had its start about as soon as Lorenz (1956) introduced PCs into meteorology (Gilman 1957), and there has been considerable discussion in the literature, especially in the climate community, as to whether physical meaning should be attached to the components (e.g., Legates 1991, 1993; Richmond 1986, 1993), but much of this relates to limited areas, neither global nor hemispheric. However, essentially the same cautions in the interpretation of the PC "patterns" should also

be heeded for the interpretation of canonical correlation analysis (CCA) patterns.

I understand HWM's method to consist of finding the largest 20 PCs for the ensemble mean week-2 forecast 500-mb height and the largest 20 PCs of the corresponding weekly mean verifying analyses. Then, canonical correlation was used to relate these two sets of 20 patterns through other patterns. Using 20 PCs to represent the forecasts and the analyses retains most, but not all, of the original variance of the forecasts and analyses. To the extent the process retains the variance, the canonical correlation can be considered equivalent to relating the original fields.

Physical interpretation of the first pair of canonical patterns—predictor and predictand (if they are used as such)—is legitimate. HWM show the predictand pattern and give the correlation as 0.81. However, for any subsequent set of patterns, one should remember that 1) the time series constructed from the predictor pattern is uncorrelated with all previous ones (this is the same constraint imposed in PC analysis), 2) the time series constructed from the predictand pattern is uncorrelated with all previous ones, and 3) all predictor time series are uncorrelated with all predictand time series, except its pair—the one yielding the nonzero correlation (note that the sign of the correlation does not matter). This is a severe restriction. One can say the *i*th predictand pattern cannot "look much like" any other pattern, even though they are not necessarily orthogonal, except possibly the *i*th predictor pattern. Generally, the first pattern is very large scale, and succeeding patterns are smaller in scale. Statements about "most predictable patterns," except the first, may be misleading when they have been produced by CCA analysis. It may be that there is a second pattern that is more predictable than the second CCA pattern, but it will produce a time series correlated with the first. Of course, a pattern that is different from the first by a minuscule amount would be about as predictable as the first, so how does one find the second and succeeding ones? CCA has required the time series to be uncorrelated. If one truly wants physically meaningful or predictable patterns, some method, such as that used by Wallace and Gutzler (1981), may be more appropriate. Principal components can be rotated to give more physically meaningful patterns (see, e.g., Richmond 1986 and Horel 1981), but to my knowledge the CCA patterns cannot without destroying their paired relationship (Wallace et al. 1992, p. 576). Nevertheless, regardless of their separate physical interpretation, the canonical functions can be used in prediction (e.g., Glahn 1968; Barnett and Preisendorfer 1987; Wallace et al. 1992).

Another comment regards the correlations of 0.6 and 0.7, between the predictor (not shown) and predictand patterns (HWM, Fig. 11), which HMW characterize as "remarkable." Large correlations are to be expected, because there are so many degrees of freedom in finding the fit. Also, a correlation of that magnitude does not necessarily imply high predictability for the following two reasons: 1) The reduction of variance, a better measure of predictability than the correlation, is only 0.36–0.49 of one pattern by the other; the relationship is statistically symmetric, and 2) each predictand pattern explains only a fraction of the total original variance. Thus, if pattern 3, for instance, explains 15% of the total variance of the predictand (not stated in HWM), a reasonable amount for such datasets (Wallace et al. 1992, p. 807, give 14% for 15 winters), the variance of the predictand over the hemisphere explained by predictor pattern 3 is only $0.15 \times 0.36 = 0.054$. That is not a lot, and its significance has not been determined. One way of thinking about this is that the time series generated by the predictor pattern does fairly well in predicting the time series stemming from the predictand pattern, but the latter pattern may not be strongly related to a real-world pattern. Table 1 in Glahn (1968) shows the relationship between the canonical correlation coefficients, the associated reductions of variance, and the actual variance of the predictand set explained for the study presented there. While the datasets are substantially different than those in HWM, the performance of canonical correlation is essentially the same. Equation (47) in that reference defines the "composite correlation coefficient" ("CCC"); the contribution to that summation by each pair of canonical functions can be considered the "partial composite correlation coefficient" ("PCCC").

To restate, HWM is a very interesting paper and may well lead to better interpretative methods. Only time will tell whether a long sample of reforecasts made at a lower resolution than the operational model will play a large role or not. Certainly, when a new model, either ensemble or not, is implemented, it would be wise to run it on a few years of data prior to implementation not only to provide a statistical sample for interpretation in terms of local weather, but to make sure it performs well over all seasons and different ENSO situations.

Perhaps this discussion will stimulate even more.

REFERENCES

Barnett, T. P., and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for

United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115,** 1825–1850.

Gilman, D. L., 1957: Empirical orthogonal functions applied to thirty-day forecasting. Massachusetts Institute of Technology, Scientific Report 3, 129 pp.

Glahn, H. R., 1968: Canonical correlation and its relationship to discriminate analysis and multiple regression. *J. Atmos. Sci.*, **25,** 23–31.

——, and D. L. Jorgensen, 1970: Climatological aspects of the Brier P-Score. *Mon. Wea. Rev.*, **98,** 136–141.

——, and D. P. Ruth, 2003: The new digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, **84,** 195–201.

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

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132,** 1434–1446.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87,** 33–46.

Horel, D., 1981: A rotated principal component analysis of the northern hemisphere 500 m height field. *Mon. Wea. Rev.*, **109,** 2080–2092.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77,** 437–472.

Legates, D. R., 1991: The effect of domain shape on principal components analysis. *Int. J. Climate*, **11,** 135–146.

——, 1993: The effect of domain shape on principal components analysis: A reply. *Int. J. Climate*, **13,** 219–228.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Massachusetts Institute of Technology Statistical Forecasting Project, Scientific Report 1, 49 pp.

Mesinger, F., and Coauthors, 2006: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*, **87,** 343–360.

Richmond, M. B., 1986: Rotation of principal components. *J. Climatol.*, **6,** 293–335.

——, 1993: Comments on: The effect of domain shape on principal components analysis. *Int. J. Climate*, **13,** 203–218.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the northern hemisphere winter. *Mon. Wea. Rev.*, **109,** 784–812.

——, C. Smith, and C. S. Bretherton, 1992: Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *J. Climate*, **5,** 561–576.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press. 627 pp.