22

The attributes of forecast systems: a general framework for the evaluation and calibration of weather forecasts

Zoltan Toth
National Centers for Environmental Prediction, Washington DC

Olivier Talagrand
Laboratoire de Météorologie Dynamique, Paris

Yuejian Zhu
National Centers for Environmental Prediction, Washington DC

Reliability and resolution are the two main attributes of forecast systems. These attributes statistically relate the performance of a forecast system to verifying data in an abstract sense. Forecast attributes have been separately defined in the literature for systems that generate forecasts of particular formats or types. In this chapter, statistical reliability and resolution are defined in a general sense, irrespective of the type or format of a forecast. Statistical reliability is concerned only with the form of forecasts, whereas statistical resolution is concerned only with the predictive capability of a forecast system, related to the time evolution of the system that is being forecast.

The two main attributes are independent characteristics of a forecast system and can be quantitatively assessed by a host of different verification measures. The general definition of forecast attributes allows a systematic discussion of the relationship between the verification and calibration of forecasts. Calibration as defined here is an adjustment of the form of the forecasts, to match the distribution of verifying observations that follow the issuance of forecasts of a particular form.

Resolution, as the inherent predictive value of forecast systems, is the attribute most sought after by developers of forecast systems. Reliability, however, is equally important in real world applications. That calls for the generation of a long enough

Predictability of Weather and Climate, ed. Tim Palmer and Renate Hagedorn. Published by Cambridge University Press.

record of hindcasts to allow for a good calibration of forecasts, or, preferably, for improvements in forecast systems that directly lead to better reliability.

22.1 Introduction

There exists a vast array of statistics for the description of various aspects of forecast systems, such as those discussed for weather and climate in this volume by Allen *et al.*, Anderson, Buizza, Hagedorn *et al.*, Kalnay *et al.*, Krishnamurti *et al.*, Lalaurette and der Grijn, Mylne, Tibaldi *et al.*, Waliser, and Webster *et al.* Some of these statistics are based solely on the forecast system investigated, while others, called verification statistics, depend both on the forecast values and the corresponding observations from the system that is being forecast (the atmosphere in the case of weather forecasts). The specifics of these statistics or forecast verification measures are not the subject of the present study. Interested readers can find a review of many of these statistics, with additional references, for example, in a recent handbook edited by Joliffe and Stephenson (2003).

Instead, this study focuses on the underlying *statistical verification attributes* of forecast systems. The main statistical forecast verification attributes, *statistical reliability* and *statistical resolution* (from here on, reliability and resolution), have long been discussed in the literature (see, for example, Murphy and Daan, 1985, and references therein). Yet these attributes have been discussed only with respect to particular forecast formats (single value, categorical, or one or another of the probabilistic forecast format types; see, e.g., Stansky *et al.*, 1989; Wilks, 1995; Joliffe and Stephenson, 2003) and not for weather forecasts of any type in general.

Sections 22.2 and 22.3 will introduce a general definition and discuss some characteristics of statistical forecast verification attributes (in short, forecast attributes), respectively. Section 22.4 will explore the statistical limits of measuring forecast attributes. Based on the general definition of the forecast attributes, and on an analysis of the statistical limitations in assessing them, an examination of the relationship between forecast verification and the calibration of weather forecasts (that is, the enhancement of certain statistical properties of the forecasts) follows in Section 22.5. Section 22.6 will explore the significance of the two main forecast attributes to developers and users of forecast systems, while Section 22.7 offers a summary of the main findings of this study.

22.2 Definition of forecast attributes

Forecast attributes, as their name suggests, are abstract concepts that the various verification statistics, using different metrics, quantify. Taking an example from physics,

length is an attribute that can be measured by a number of different metrics. As mentioned in the Introduction, forecast attributes have been discussed so far in the context of specific types of forecasts (see, e.g., Murphy and Daan, 1885; Stansky *et al.*, 1989; Wilks, 1995; Toth *et al.*, 2003). Forecast attributes are defined below in a general sense, allowing for a comprehensive discussion of weather forecasts and their statistical calibration.

The verification attributes discussed below are defined in a statistical sense, which is related to forecast systems, and not to individual forecasts generated by them. Forecasts can be of any format but are assumed to belong to a finite number of different 'classes', called F_i. The set of verifying observations corresponding to a large number of forecasts of the same class are characterised by an empirical frequency distribution, called observed frequency distribution (ofd), and marked by o_i.

22.2.1 Reliability

When defining the first forecast attribute, statistical reliability, consider a particular forecast class, F_i. Consider further the frequency distribution of observed outcomes that follow forecasts from class F_i, that is o_i. If forecast F_i has the exact form of o_i for all forecast classes (i), the forecasts are statistically consistent with the observations and the forecast system is called (perfectly) reliable. Different measures of reliability are based on various methods for comparing forecast F_i and the corresponding observed frequency distribution o_i for all forecast classes (i), and measuring their difference.

22.2.2 Resolution

The second forecast attribute, statistical resolution, is defined as a forecast system's ability to distinguish, ahead of time, between different outcomes of the natural system (in case of weather forecasts, the future state of the real atmosphere).

For a more formal definition of resolution, let us assume that the observed events are classified into a finite number of classes, marked by O_i. If each observed class O_i is preceded by a distinctly different forecast class F_i, the forecast system is said to have perfect resolution. Conversely, if the forecast is the same prior to each observed class O_i (i.e., the forecasts do not vary, $F_i = F$ for all i), or if the forecasts vary but the observed frequency distribution o_i following the issuance of different forecasts F_i is the same (i.e., $o_i = c$, the climatological distribution, for all i), the forecast system has no resolution at all.

Resolution in a forecast system can be measured by the degree of separation among the frequency distributions of observed events (o_i), conditioned on different

forecast classes (F_i). In practice, this can be achieved by comparing the observed frequency distributions (o_i), constructed from observed events that follow different forecast classes, with the overall climatological distribution of observations (c, that is the reference for a forecast system with no resolution). Different measures of resolution are based on various methods for carrying out this comparison.

22.3 Some characteristics of forecast attributes

(a) *Reliability and resolution are two independent attributes*. Reliability is concerned only with the statistical consistency between each class of forecasts F_i and the corresponding distribution of observations o_i that follow such forecasts, whereas resolution is not affected at all by this consistency. By contrast, resolution reflects how well different forecast classes can separate cases with different subsequent observed events, whereas reliability is unaffected by this property of forecast systems.

While the format and the actual values used by a forecast system are irrelevant to its resolution, they are critical for its reliability. By contrast, a forecast system with perfect reliability does not necessarily have good resolution. Two examples are interesting to note here. A forecast system always issuing the observed climatological distribution has perfect reliability and no resolution by definition, while a system using forecast anomalies that are systematically reversed compared with observed anomalies would have perfect resolution but no reliability.

(b) *In principle, reliability can always be statistically 'enforced' or corrected.* This is true as long as both the forecast and observed systems are stationary in time, and there is a long enough record of forecast-observed data pairs. This is because reliability reflects only the statistical consistency between forecast and observed distributions. All one has to do to achieve the desired consistency is to replace the forecasts in a given forecast class with the frequency distribution of observations that follow such forecasts.

(c) *Unlike reliability, resolution cannot be improved by statistically correcting the forecasts so they follow the distribution of ensuing verifying observations*. This is because resolution does not depend on statistical consistency. Resolution reflects the inherent value of forecast systems, and can be improved only through the modification of the forecast scheme based on additional knowledge about the temporal evolution of the observed system.

(d) *Reliability and resolution, as defined above, are general attributes of forecast systems*. They can be interpreted for systems generating forecasts of any type, such as single value, categorical, or probabilistic.

It is interesting to note that single value (out of a continuum) forecasts can be perfectly reliable only if they have perfect resolution as well. This is the only way the observed frequency distribution would exactly match the Dirac function form of the forecasts.

As mentioned earlier, forecast attributes have been interpreted in the past for forecasts issued in specific formats (i.e. not necessarily in the general form of a probability distribution). While this can be useful for special purposes, it must be noted that such narrow definitions of forecast attributes are not fully consistent with the general definition introduced in this study.

Consider, for example, the case of a forecast system with less than perfect resolution that issues single value forecasts. In this case, it could be possible to define statistical reliability (or statistical consistency, as it is also often referred to; see, e.g., Wilks, 1995) as a lack of conditional systematic bias. According to this narrow definition, a forecast system is considered reliable if for all forecast values the frequency distribution of corresponding observations has the same mean as the forecast value. It is easy to see that this feature is a necessary but not sufficient condition for reliability as defined in the present study. In fact, the no-spread single value forecasts, even if they have no systematic bias, will have less than perfect reliability for any system with less than perfect resolution. Such a narrow definition of reliability will have an implication for statistical calibration as well, as it will be discussed in Section 22.5.

22.4 The limits of assessing reliability and resolution

22.4.1 Measures of forecast attributes

As discussed by Toth *et al.* (2003) for forecasts in probabilistic format, some existing verification measures assess reliability, some resolution, while still others provide a combined measure of both. Note that some measures can be calculated for selected subsets of all forecast cases – like the reliability and resolution components of the Brier score verifying for only one of a set of categorical events. These measures can be related to reliability and resolution as defined in the present study only if the measure is aggregated over all observed categories.

22.4.2 Factors limiting the statistical accuracy of verification statistics

While forecast attributes can theoretically be defined assuming that the number of forecast cases goes to infinity, in practice verification measures are always computed based on finite samples. Therefore, verification results can be considered estimates whose accuracy will depend on the sample size. Knowledge about the uncertainty in verification results is important (see, e.g., Hamill, 1997), especially when one

compares two or more competing forecast systems. In such cases it is especially important to assess the statistical significance of the comparative verification results (see, e.g., Candille and Talagrand, 2005). The associated uncertainty in the verification results can be reduced only through increasing the sample size, which is often impossible when evaluating real life forecast systems.

Another factor limiting the accuracy of verification estimates is the uncertainty in the verifying data (Candille, 2003). Observations used to verify forecasts are generally associated with measurement and other errors. For properly assessing reliability and resolution of a forecast system, such errors in the observations need to be carefully accounted for, otherwise the results will either be biased and/or will look statistically more certain than they are. Observational errors can be considered in forecast verification by replacing an observed value (Delta function) with a probability density function (pdf) that reflects the observational uncertainty. The use of incorrect observational error estimates (such as assuming perfect observations in the presence of errors, as in the case of most verification studies) will introduce errors in the verification (and pursuant calibration) results.

A third factor influencing the accuracy of forecast verification statistics is the choice of the level of granularity introduced in the calculations, which is a function of the level of detail sought in the results. The granularity of verification studies can be controlled through a number of choices.

First, forecasts can theoretically take an infinite number of forms. Yet, when in practice a finite sample of forecasts are evaluated statistically, forecasts of a similar form must be grouped into a finite number of classes. For more detailed verification statistics one might possibly wish to establish a large number of forecast classes. The number of different classes is limited, however, by the requirement that there be enough forecast cases in each of the classes established.

Second, forecast probability distributions can theoretically be defined and manipulated as continuous functions. In practice, however, calculations are always carried out over finite intervals. And because the sample size is limited, the width of the intervals cannot be reduced arbitrarily, otherwise most intervals would contain no data points.

Finally, if the overall sample size is small, one may need to group together forecast–observed pairs from similar geographical regions and/or similar parts of the annual cycle.

In practice, when choosing the level of granularity in verification calculations, one seeks a compromise between having a large enough sample for all forecast classes and verification intervals, while retaining as many classes, intervals, and geographical, seasonal distinctions as possible, given the total number of forecast–observation pairs (Atger, 2003). Obviously, the larger the overall sample of forecast–observation pairs is for verification, the more questions about the performance of the forecast system can be answered. As we will see in the next section, the same holds true for the number of adjustment types that can be made as part of a statistical calibration algorithm.

22.5 Calibration

The goal of calibration is to make the form for each class of forecasts statistically more consistent with the distribution of the corresponding verifying observations. Calibration, as defined here, is the replacement of the forecast, whatever form it may have (i.e. single value, categorical, or probabilistic), with an estimate of the corresponding odf (which describes the distribution of observations that in the past followed the issuance of forecasts from the same forecast class). The success of calibration can be measured by comparing the reliability of the calibrated forecasts with that of the raw, uncalibrated forecasts.

Note that calibration is directly related to the verification of statistical reliability, since both are based on estimating the distribution of observations following different forecast classes. While verification assesses the statistical reliability of a forecast system over a period in the past, calibration adjusts the forecasts with the intention to make them more consistent with observed statistics in the future. Calibration is based on the assumption that the statistical behaviour of the forecast and observed systems, as analysed over a period in the past, will not change in the future. Calibration, therefore, is subject to an additional limitation beyond those discussed with respect to verification, namely that the quality of calibration will suffer if either the natural or the forecast system is non-stationary in time. As with verification, small sample size, the presence of uncertainty, and errors in describing uncertainty in the verifying observations will also adversely affect calibration results, as will an inappropriate choice for the level of granularity in the calculations.

There are a number of ways that forecasts from different classes, geographical regions or different parts of the annual cycle can be grouped together for computing verification statistics that are also needed for calibration. The resulting formation of larger subsamples allows a more robust statistical estimate of the underlying distribution of the observations corresponding to a broader group of forecasts – at the expense of reducing the level of details in the verification, and consequently in the pursuant calibration results. Therefore, careful compromises are needed when the level of granularity is chosen for the computation of statistics for calibration. Allow too many details in the verification (i.e. use too many different forecast classes), and the calibration will suffer from sampling noise. Conversely, the lack of enough detail in verification (i.e. grouping forecasts from areas with distinctly different verification statistics together; see Atger, 2003) can also adversely affect the calibration by leaving the biases present in the smaller subsamples uncorrected.

It should be noted that calibration, as discussed earlier with respect to verification, can be introduced in a narrower sense than that defined above. Forecasts, for example, can be corrected only to reduce their systematic bias in the first moment. An application based on such a narrow definition of calibration will necessarily be limited since other, higher moment aspects of the forecasts will not be statistically

corrected. By contrast, calibration, if applied in a general sense as defined above on single value, categorical, or any other type of forecasts, will naturally change the format of the forecasts to the more general probabilistic format.

22.6 Significance of attributes to forecast developers and users

Neither the reliability nor the resolution of real life weather forecast systems is perfect. What is the significance of either attribute to the developers or users of weather forecasts? Is one or the other attribute more important?

22.6.1 Developers' perspective

We recall that the inherent value of forecast systems lies in their ability to predict future events, as reflected in the statistical resolution of forecast systems. This is equivalent to a forecast system issuing uniquely different signals prior to different observed events. For example, if a system systematically gives a prediction of 'heavy snow' (or 'red') and 'light snow' (or 'blue') prior to observed rain and no rain events respectively, it has a high resolution.

Since the forecast signals issued by this forecast system are significantly different from the subsequent observed verification events, however, the forecasts have poor reliability. If such behaviour is systematic, the forecasts can be calibrated and the developers of the forecast system may be content with the good resolution and may not be overly concerned with the apparent lack of reliability.

22.6.2 Users' perspective

It must be noted that when forecasts from the system described above are taken by the users at their 'face value', they can be worthless or even harmful. A user who believes what the forecast says and acts on that information can be seriously hurt (e.g. Zhu *et al.*, 2002). Even forecast systems with high predictive skill (high resolution) have no value to users unless they also have good reliability. This explains why users often emphasise reliability in their evaluation of forecast systems, based on the principle of 'do no harm'.

22.6.3 Need for calibration

Generally, a long enough record of observed–forecast pairs will allow an adjustment or calibration of the forecast signal to match the distribution of observations that follow a particular forecast class. Incidentally, a similarly long record of observed-forecast pairs may be needed for the precise assessment of resolution in a forecast

system (see Section 22.4). In the case of a forecast system with high resolution, calibration can significantly enhance the utility of forecast systems. This underlies the need for the provision of a large enough set of hindcasts (forecasts generated on past events). This will allow a proper assessment of both the resolution and reliability of the forecast system, and will facilitate a subsequent calibration of the forecasts in case the forecast system lacks statistical reliability. In such a case, statistical reliability can be achieved through a statistical adjustment via calibration.

22.6.4 Value of forecasts

As discussed above, beyond resolution, the users also critically depend on the reliability of the forecasts. It is therefore important that when (typically after they are calibrated) the value of forecast systems is assessed for the users, both resolution and reliability are considered.[1] One can argue that for a forecast system to show genuine improvement, its resolution must be measurably enhanced. An experimental forecast system with enhanced resolution, but an insufficient hindcast data set for calibration, however, may degrade utility. One may argue that enhanced resolution forecast systems be operationally implemented only if their reliability is not affected negatively, or if at least a sufficient hindcast dataset is generated to ameliorate the problem through calibration.

22.6.5 Future directions

As forecast systems mature, there is a natural tendency to use more detail from the forecasts. For that to happen, one needs to include more detail in the calibration of the forecasts as well. That, as discussed earlier, calls in turn for longer periods of past observed–forecast pairs. Unfortunately, the number of such pairs is usually severely limited due to the lack of long periods of detailed observations. This is of particular concern when extreme events are considered. Such events, by definition, occur rarely (Zhu and Toth, 2001). Therefore, their statistical calibration is especially problematic (Legg and Mylne, 2005). Yet these rare events are often of the greatest interest to users.

It follows that as forecast and application methods improve and more details are demanded from a system, the potential value added by statistical calibration will likely diminish. Since under such conditions statistical corrections are of little or no help, directly improving the reliability of a forecast system itself will become more important and sometimes will offer the only tractable solution. When the realism of models representing weather systems (that is directly related to reliability) is improved, the changes may also lead to improvements in predictive skill (i.e. resolution). Prediction of tropical storms is a prime example of a situation where the role of statistical calibration is limited due to the highly non-linear nature of these systems. If a storm, due to model deficiencies (e.g. too low spatial resolution), is not

predicted (well) by a forecast system, the insertion (modification) of a storm into the forecast via statistical inference/calibration may require an impractically large training data set. In such cases the reliability (and utility) of the forecasts can be improved only by enhancing the realism of the numerical weather prediction model itself.

22.7 Conclusions

This study introduced a distinction between the abstract notion of forecast system attributes and the statistical measures used to assess them. Unlike earlier studies, a general definition of the forecast attributes was proposed, irrespective of the format of the forecasts. Both of the two main attributes, reliability and resolution, were interpreted in a statistical sense. Reliability was defined as a perfect match between the form of a forecast and the distribution of verifying observations that follow the issuance of that particular forecast form. A forecast system is said to have perfect resolution, by contrast, if it consistently gives different signals prior to the occurrence of different observations.

Reliability and resolution were shown to be independent of each other. Of the two attributes, forecast system developers are more concerned about resolution since that is related to the intrinsic predictive capability of forecast systems. For the users who take the weather forecasts at face value, reliability is equally or even more important. This is because it is reliability that assesses how what is being forecast (i.e. the form of the forecasts) and directly acted upon by the users compares statistically with what is being observed.

A number of verification measures exist for the assessment of reliability and resolution. These measures, like any other statistics based on finite samples, are subject to sampling and other types of errors. These same errors were also shown to affect calibration, where the reliability of forecast systems is enhanced. Calibration was defined in general terms as the replacement of the form of the forecasts by the distribution of observations that follow the issuance of any particular forecast form, based on a set of observed–forecast data pairs.

It follows from the general definition of the main forecast attributes and calibration that the general format of forecasts is that of a probability density function (pdf) since that is the only format that can, in general, be consistent with the distribution of ensuing observations. A pdf format allows the forecast system to reflect case-by-case variations not only in the expected first moment of future weather parameters but also in the higher moments, such as error variance. For example, forecasts in pdf format can distinguish, given a certain expected value, between cases with higher and lower uncertainty (Toth *et al.*, 2001). Such information is known to have potentially great economic value for the users (Zhu *et al.*, 2002), yet cannot be provided by a forecast system using a single value format. To what extent ensemble forecast systems can

provide useful information beyond the first moment of the distribution is still an open question (see, e.g., Atger, 1999).

Acknowledgements

This chapter is an outgrowth of work on a chapter (Toth *et al.*, 2003) written by the authors and a collaborator (G. Candille) for a textbook edited by Joliffe and Stephenson (2003). The authors acknowledge the stimulating discussions with the editors of that volume. The first author is indebted to Professor Eugenia Kalnay, who asked him to contribute to a lecture series on Statistics in Meteorology at the Department of Meteorology, University of Maryland. This chapter is also intended as a draft contribution to the lecture notes accompanying that course.

Note

1. As discussed in Zhu *et al.* (2002), some measures of forecast performance, such as the potential economic value, assume that the forecasts can be perfectly calibrated (i.e. forecasts are automatically calibrated as part of the computation of potential economic value, using the *dependent* and not an independent set of data for calibration). These results will overestimate the actual utility of forecasts that in practice will necessarily be lowered by the limits of calibration discussed in Section 22.5.

References

Atger, F. (1999). The skill of ensemble prediction systems. *Mon. Weather Rev.*, **127**, 1941–53.

Atger, F. (2003). Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Weather Rev.*, **131**, 1509–23.

Candille, G. (2003). *Validation of probabilistic meteorological forecast (in French)*. Doctoral Dissertation, Universite Pierre-et-Marie-Curie, Paris, France.

Candille, G. and O. Talagrand (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–50.

Hamill, T. (1997). Reliability diagrams for multicategory probabilistic forecasts. *Weather Forecast.*, **12**, 736–41.

Joliffe, I. T. and D. B. Stephenson (eds.). (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley.

Legg, T. P. and K. Mylne (2005). Early warnings of severe weather from ensemble forecast information. *Weather Forecast.*, **19**, 891–906.

Murphy, A. and H. Daan (1985). Forecast evaluation. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, ed. A. H. Murphy and R. W. Katz, pp. 379–437. Westview Press.

Stanski, H. R., L. J. Wilson and W. R. Burrows (1989). *Survey of Common Verification Methods in Meteorology*. World Weather Watch Technical Report 8. World Meteorological Organization.

Toth, Z., O. Talagrand, G. Candille and Y. Zhu (2003). Probability and ensemble forecasts. In *Forecast verification: A Practitioner's Guide in Atmospheric Science*, ed. I. T. Jolliffe and D. B. Stephenson, pp. 137–63. Wiley.

Toth, Z., Y. Zhu and T. Marchok (2001). On the ability of ensembles to distinguish between forecasts with small and large uncertainty. *Weather Forecast.*, **16**, 436–77.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*. Academic Press.

Zhu, Y. and Z. Toth (2001). Extreme weather events and their probabilistic prediction by the NCEP Ensemble Forecast System. In *Preprints of the AMS Symposium on Precipitation Extremes: Prediction, Impact, and Responses*, *14–19 January 2001, Albuquerque, NM*, pp. 82–5. American Meteorological Society.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson and K. Mylne (2002). The economic value of ensemble based weather forecasts. *Bull. Am. Meteorol. Soc.*, **83**, 73–83.