CHAPTER 21. NUMERICAL PREDICTION OF THE EARTH SYSTEM: CROSS-CUTTING RESEARCH ON VERIFICATION TECHNIQUES

Elizabeth Ebert, Barbara Brown, Jing Chen, Caio Coelho, Manfred Dorninger, Martin Göber, Thomas Haiden, Marion Mittermaier, Pertti Nurmi, Laurence Wilson and Yuejian Zhu

Abstract

In the last decade or so the meteorological community has seen the successful development and application of new and improved forecast verification methods for numerical prediction of the Earth system. Verification methods to evaluate ensemble forecasts have become essential because of the prominent role of Ensemble Prediction Systems as sources of numerical guidance in operational centres. Moreover, coupled atmosphere-land-ocean models are now routinely run in major centres to provide operational predictions on seasonal and multi-week time frames. This chapter focus on advances in methods for evaluation of forecasts of several different types of phenomena, as well as methods for different types of forecasts and different timescales.

21.1 INTRODUCTION

Numerical Weather Prediction (NWP) model forecasts have been verified since the 1950s when they first started providing reasonable predictions. Several World Meteorological Organization (WMO) Lead Centres for forecast verification[a] now coordinate the routine production of verification results for NWP and seasonal climate predictions from major national centres. Moreover, WMO's Commission for Basic Systems (CBS) encourages the exchange of standard verification scores for NWP models. In the case of NWP, these score included bias, root mean square error (RMSE), S1 skill score, and anomaly correlation of forecast fields on selected pressure levels. Recently, verification scores for surface parameters have been added to the exchange in recognition that the accuracy of surface parameter forecasts has improved as a consequence of scientific and technical advances in NWP capabilities and increased horizontal spatial resolution of many global models, which is below 20 km in many cases.

Verification methods to evaluate ensemble forecasts have become essential because of the prominent role of Ensemble Prediction Systems (EPSs) as sources of numerical guidance in operational centres. Moreover, coupled atmosphere-land-ocean models are now routinely run in major centres to provide operational predictions on seasonal and multi-week time frames. The benefits of this coupling at shorter ranges have also been recognized, and it is being applied by some centres; for example, the European Centre for Medium Range Weather Forecasting (ECMWF) EPS is now coupled from day 1. NWP and climate models are routinely used to drive downstream impact models for emergency management, hydrology, agriculture, energy, and many other applications. These downstream developments highlight the need for users to be involved in the evaluation process.

This volume describes the many advances made in numerical prediction and the challenges to be addressed in coming years. Improvements in numerical prediction require improved methods to verify these forecasts. This has been an active area of research in the last decade or two. The World Weather Research Programme (WWRP)/Working Group on Numerical Experimentation (WGNE) Joint Working Group on Forecast Verification Research (JWGFVR) was established in 2003 to promote work in this area. This group coordinates workshops, tutorials, and verification method intercomparisons, and is the focal point for verification of WWRP Forecast and Research Demonstration Projects (FDPs and RDPs).

[a] *The Lead Centre for Deterministic Forecast Verification is located at the European Centre for Medium Range Weather Forecasts (ECMWF), the Lead Centre for Ensemble Forecast Verification is located at the Japan Meteorological Agency (JMA), and the Lead Centres for Long Range Forecast Verification are located at the Australian Bureau of Meteorology and the Meteorological Service of Canada.*

This paper describes some of the recent successes as well as current challenges facing the verification community, as reflected in recent workshops and other presentations and papers. The following sections focus on advances in methods for evaluation of forecasts of several different types of phenomena, as well as methods for different types of forecasts and different time scales. Remaining research and challenges associated with each of these topics are also considered. The final section summarizes the current state-of-the-art in forecast verification and describes some additional challenges in verification research and applications.

21.2 SPATIAL VERIFICATION METHODS

Short and medium-range NWP models have improved considerably over the years, while they have also been evolving toward ever-higher resolution. Moreover, the prediction of surface weather parameters has greatly improved. The spatial variability and intensity distributions of model variables increasingly resemble the variability and distributions of observations, and the ability to simulate the extreme values that are very important in a forecast and warning context is also improving. Traditional verification against standard observations may suggest that high resolution forecasts are less accurate than the lower resolution ones (e.g. Mass et al. 2002). The spatial and temporal scales of the verification have a strong influence on the measured performance with finer scales more prone to the "double penalty" associated with small errors in the location and intensity of a forecast feature.

To measure the performance of high resolution forecasts in a way that is more consistent with how they are used, several new spatial verification approaches have been proposed. Gilleland et al. (2010) describe these methods as neighbourhood (crediting "closeness" in space, time, and/or intensity, often through probabilistic approaches), scale separation (quantifying error at various scales), features-based (comparing attributes of forecast and observed weather features such as their location, size, intensity, etc.), and field deformation (measuring the distortion required to make the forecast resemble the observed field). Gilleland et al. (2010) compare more than a dozen spatial verification methods and their respective ability to measure location, intensity, and structure errors, distinguish skilful scales, and verify the predicted occurrence of events. Table 1 gives a summary of these capabilities by the type of verification approach. While all methods measure intensity bias, no single method addresses all types of errors. Therefore, it is necessary to either prioritise which types of errors are most important to the user and choose the appropriate verification approach, or preferably apply more than one type of verification method. More complex verification methods could be developed that address a greater range of error types.

Table 1. Intercomparison of traditional and spatial verification methods (after Gilleland et al. 2010). A tick indicates that the method addresses the given type of error, a cross indicates that it does not.

Category	*Scales with skill*	*Location errors*	*Intensity errors*	*Structure errors*	*Occurrence (hits, misses, false alarms)*
Traditional (gridpoint)	×	×	✓	×	✓
Neighbourhood	✓	×	✓	×	✓
Scale separation	✓	×	✓	×	✓
Features based	×	✓	✓	✓	✓
Field deformation	×	✓	✓	×	×

Neighbourhood and feature-based methods are becoming mature enough to be used routinely in many NWP centres to verify high resolution models. For example, the Met Office uses the fractions skill score (FSS; Roberts and Lean, 2008) and neighbourhood Brier score (Mittermaier, 2014a) to measure the scales at which the model shows useful skill for predicting rainfall and clouds. The Method for Object-based Diagnostic Evaluation (MODE) and the Contiguous Rain Area (CRA) method are used to characterize performance of rainfall forecasts in many national centres (Ebert and McBride, 2000; Davis et al. 2009, http:/www.hpc.ncep.noaa.gov/verification/mode/mode.php - page=page-1. These methods can be applied to other parameters (for example, wind, moisture, cloud) and can evaluate timing errors, though more research is needed to explore these possibilities further.

Advanced verification methodologies have mostly focused on rainfall, due in large part to the availability of spatially and temporally complete quantitative precipitation estimates from radar. These methodologies must be tested for their ability to provide useful performance information for other variables such as wind, waves, pollutants and other hazards, as well as for more benign variables like temperature, humidity and cloud cover. The second spatial verification method intercomparison project, called MesoVICT (Mesoscale Verification In Complex Terrain; Dorninger et al. 2013) focuses on testing spatial verification methods on precipitation and wind forecasts from both deterministic and ensemble NWP, and for the first time includes ensemble analyses as reference data to simulate uncertainty associated with observation fields. Verification researchers are encouraged to participate in this project to test existing and newly developed spatial verification approaches and to explore how to account for observational uncertainty.

In light of growing reliance on high resolution NWP models for predicting high-impact weather, work must continue on characterizing location and intensity forecast errors for small scale intense features - features for which small errors can have large impacts on decision-making and impacts. It may be possible in some circumstances to correct for systematic forecast biases, particularly where intensity biases are related to model resolution. Moreover, with short-range forecasts now being used to produce graphical (spatial) warnings, there is a strong need to extend spatial verification methods to evaluate timing (lead time, onset, cessation) and intensity performance of spatial warnings.

Spatial verification approaches also have potential to provide valuable insights on the performance of longer lead time (multi-week to seasonal) forecasts through their use to evaluate anomaly predictions (e.g. for temperature and precipitation). Neighbourhood methods are used to assess the value of downscaling (De Haan et al. 2014), while features-based approaches can be used to characterize errors in anomaly or other patterns (e.g. mean, variance, extreme fields) to provide more intuitive information for users and service providers. This topic is a current area of research.

21.3 METHODS FOR EXTREME EVENTS

A strong motivation for high resolution NWP is to predict extreme values associated with dangerous weather. One challenge in verifying predictions of extremes is the limited frequency of opportunities to observe them and, thus, collecting enough forecast-observation pairs to compute meaningful and robust statistics. Thus, the accuracy of extreme event forecasts can be challenging to assess. Moreover, in cases of extreme weather the observations themselves may be less trustworthy; for example, instruments may be destroyed by floods or windstorms or measurements may be compromised by the weather conditions.

Some common categorical contingency table-based verification metrics behave badly for rare events, making them ineffective at distinguishing variations in performance among multiple forecasting systems. In particular, Ferro and Stephenson (2011) show that for imperfect forecasts, the threat score and the Gilbert, Heidke, and Peirce skill scores asymptote to zero for rare events. Many of these scores are strongly affected by the number of correct non-events and provide little useful information on the model performance for rare events. Ferro and Stephenson have proposed a new class of scores called extremal dependence scores (EDSs) that reward hits and penalize misses and false alarms, and also behave much more consistently with the forecast performance observed for less rare events. The simplest EDS is the extremal dependence index (EDI), defined as

$$EDI = \frac{(\ln F - \ln H)}{(\ln F + \ln H)},$$

where *F* is the false alarm rate and *H* is the hit rate (both of which must be non-zero). While the interpretation of the EDI is less clear cut than for the threat score, it has the strong advantage of being able to better distinguish the performance of competing models for rare binary events. Extreme value theory, widely used to analyze extremes in the climate context, offers some promise for evaluating the performance of extreme weather forecasts of continuous variables (Prates and Buizza, 2011). A threshold-weighted continuous ranked probability score (CRPS) was recently proposed by Gneiting and Ranjan (2011) as a strictly proper score for evaluating probability forecasts for extremes.

Forecasters increasingly rely on guidance from numerical predictions to issue watches and warnings for severe and high-impact weather. In contrast to routine forecast verification with fixed base times and valid times, warning verification requires the evaluation of lead time and warning duration relative to the onset and cessation of the event being warned for. Trade-offs between lead time and warning accuracy need to be assessed in order to inform user-focused studies of warning effectiveness in the face of false alarms (Wilson and Giles, 2013).

The spatial extent of a warning also influences the verification; it is easier to accurately warn for an event somewhere within a large area (e.g. a county or state) than in a small area like a town. Similar to neighbourhood verification, it may be desirable to apply "soft" criteria to warning verification - within *X* km, within *Y* minutes, within ±1 intensity category, etc., as well as "hard" criteria, to understand better the warning performance as function of scale and other factors (Neal et al. 2014). This is particularly true when observations are incomplete, as in the case of tornado sightings, in which case it may be necessary to treat observations in a probabilistic manner (Brooks et al. 1998; Hitchens et al. 2013).

With improving forecasts and warnings for extreme weather and its impacts becoming increasingly the focus of operational meteorology, greater efforts must be made to develop methods for objectively evaluating and communicating their performance in terms that users can understand. Relevant metrics might measure lead time, accuracy of predictions for "unsafe" and "all-clear" conditions, and dollars saved or losses averted relative to the no-forecast case or some other standard. Section 21.7 discusses user-oriented verification in more detail. A useful metric in this context is the relative (or potential) economic value (Richardson, 2000), which translates forecast skill into potential economic gain for users with different cost/loss ratios. It has been also found to be a useful framework for assessing the benefit of probabilistic vs. deterministic forecasts for extreme events (Haiden et al. 2014; Magnussen et al. 2014).

21.4 METHODS FOR ENSEMBLE AND PROBABILISTIC PREDICTIONS

Ensemble prediction is now being used at all scales to explicitly account for forecast uncertainty related to initial conditions and model uncertainties. EPSs can be evaluated in several different ways with the choice of approach dependent on how the forecast is intended to be used. Specifically, the ensemble members can be evaluated individually as deterministic forecasts or the ensemble can be summarized using a representative member such as the ensemble mean; they can be evaluated as probabilistic forecasts (e.g. by translating the ensemble prediction to a probability distribution or by estimating probabilities for specific events); or they can be evaluated as a distribution. While methods focused on the first two options are relatively well-established, methods for evaluation of a whole distribution are still relatively new, and improved diagnostic and intuitive approaches for evaluation of EPSs are still needed.

Traditional verification of probabilistic and distribution forecasts from EPSs is based primarily on metrics such as the Brier skill score (for probability forecasts) and CRPS (for the whole distribution), and diagnostics such as reliability and relative operating characteristic (ROC) diagrams and rank histograms, to assess spread-error consistency and reliability and discrimination of probability and ensemble forecasts. The ignorance score (Roulston and Smith, 2002) is also becoming more commonly applied as a single number to evaluate the quality of an ensemble distribution without inferring the preferences of a particular user. This measure has recently been decomposed to represent attributes that are similar to the attributes of reliability and resolution represented by the Brier Score (Weijs et al. 2010; Tödter and Ahrens, 2012).

The "spread-skill" relationship is often relied upon to determine the adequacy of the ensemble in appropriately capturing the forecast uncertainty; yet the methods for doing so are varied and the interpretations often not completely clear. A proposed new error-spread score (ES) verifies the moments of the forecast and is able to distinguish between dynamically reliable forecasts from an ensemble prediction system and the statistically reliable (but non-varying) dressed deterministic forecasts (Christensen et al. 2015). Hopson (2014) explores the nuances associated with different approaches to estimating and comparing spread and skill.

Ferro (2014) considers the "fairness" of scores such as the Brier Score and CRPS for evaluation of ensemble predictions. In Ferro's study a score is defined to be fair if "the expectation of the score with respect to the distributions of both the ensemble members and the verifying observation is optimized when these distributions coincide". Ferro's work indicates that the Brier, ranked probability and continuous ranked probability scores are unfair. However, appropriately adjusted versions of these scores are fair. Meta-studies like this - related to properties of scores - provide valuable guidance for the application of the scores and the interpretation of the results. They are important contributions to the verification knowledge base as new scores gain wider use.

Spatial methods are now starting to be used to evaluate ensemble predictions - this is a promising area of research and application, especially as convection-permitting ensembles become a routine tool for high-impact weather prediction in some national centres. Neighbourhood verification methods are easily extended to include an ensemble dimension (e.g. Duc et al. 2013; Ben Bouallègue and Theis, 2014; Mittermaier, 2014a), and scores such as the FSS can be used to characterize the ensemble spread (Dey et al. 2014). Approaches for feature-based ensemble verification are still being investigated (e.g. Gallus, 2010; Johnson et al. 2013). Suggested approaches include verifying objects in probability maps, verifying the "ensemble mean" using spatially averaged forecast objects (possibly with histogram recalibration) or objects generated from average object properties, and evaluating distributions of object properties.

Because the field of ensemble verification is still relatively young, there is a continuing need for more intuitive and informative methods. This will be an important area of research in the future.

21.5 UNCERTAINTY IN VERIFICATION RESULTS

Uncertainty in verification results arises from many sources. Perhaps most importantly, observations are inherently uncertain due to measurement as well as spatial and temporal representativeness errors, and application of forecast verification to limited samples of forecasts leads to uncertainty related to sampling variability. Sampling variability is somewhat more straightforward to account for than observation-related uncertainty, and methods for estimating statistical confidence intervals have been defined for many verification measures (e.g. Jolliffe, 2007; Gilleland, 2010) and are included in at least some verification packages (e.g. the Model Evaluation Tools (MET): http://www.dtcenter.org/met/users/). These approaches generally take into account the effects of temporal correlations; accounting for the impacts of spatial correlations on the confidence intervals is somewhat more problematic and is generally not adequately addressed. Methods for applying confidence intervals to differences in performance for paired samples lead to more powerful statistical comparisons of model forecast performance.

While taking into account observation uncertainty in verification studies is still a research topic, some knowledge has been gained in recent years. However, much more knowledge and new capabilities are required. Fundamentally, as models have improved, it is no longer appropriate to ignore observation error; in fact, as models improve, the apparent error in forecasts will become closer and closer to the error in the observations. Ideally, biases in observations can be removed (when they are known) but it is more difficult to account for the random errors, which lead to poorer verification scores for deterministic forecasts. Verification results for ensemble forecasts are characterised by poorer reliability and ROC area in the presence of observation error.

A few solutions have been suggested for accounting for observation error in verification analyses to try to estimate the "true" forecast performance against perfect observations. A simple example is to include error bars in scatterplots of forecasts vs. observations. Ciach and Krakewski (1999) proposed approaches for coping with observation errors in computation of RMSE values; and Bowler (2008) considered how to incorporate observation uncertainty into categorical scores. In addition, Santos and Ghelli (2011) have looked at a version of the Brier Skill Score that accounts for observation uncertainty. A difficulty with these approaches is that in the absence of a "gold standard" of the true value of the observed parameter, the observation errors are themselves only estimates and can lead to unrealistic estimates of forecast error. Triple collocation analysis can potentially provide estimates of error variances for three or more products that retrieve or estimate the same geophysical variable using mutually independent methods; however, a recent study suggests that cross-correlation of errors causes the true random error to be underestimated (Yilmaz and Crow, 2014). Mittermaier (2014b) explored the impact of temporal sampling on the representativeness of hourly synoptic observations by considering 1 minute surface observations of temperature. Though information on the variance of the residuals can be derived, it is less clear how these should be applied, and this is an area of current research. Initial results would suggest that there may be a limit to achievable forecast accuracy.

Another area of concern is the difference in forecast performance that is apparent when comparisons are made with multiple observation sources (e.g. different analyses; gauges vs. radars) (Tollerud et al. 2014). Accounting for this variability is difficult but important. Differences in analyses provide another representation of the uncertainty associated with observations and their appropriateness for matching to specific forecasts. Gorgas and Dorninger (2012) investigated the use of an ensemble of objective analyses as verification for NWP forecasts of surface variables, to quantify the uncertainty in verification results associated with the spatial treatment of the observations. The MesoVICT project will provide an opportunity to test the sensitivity of different traditional and spatial verification methods to choice of analysis, and ways to potentially exploit this analysis variability to provide useful insights on forecast performance (see Section 21.2 for more on this project).

The wisdom of using a model's own analyses (i.e. the model state at its initial time, following data assimilation) to verify its forecasts has come under increasing scrutiny following recent findings that such results may lead to incorrect conclusions about the nature of model errors (Yamaguchi et al. 2014). Model biases carry over into the analysis from the model-based background field, especially where observations are sparse, leading to underestimates of model error and over-estimates of ensemble dispersion in the short range. Members of WGNE investigated this problem, verifying NWP models from each centre against its own and others' analyses, and found some surprising behaviours including model errors sometimes improving with lead time when verified against other centres' analyses (WGNE, 2014). As a result WGNE recommended putting greater emphasis on verification against observations.

When spatially complete observation fields are required for verification, model-independent analyses such as the Vienna Enhanced Resolution Analysis (VERA; Steinacker et al. 2000) may be used confidently where the observation density supports regular gridding. In regions where observation density is highly variable, such as Canada, it may be possible to use a modified grid. Casati et al. (2014) proposed a wavelet-based objective analysis scheme in which the size of each grid box varies according to observation support; verification is then performed at different scales according to the observation availability. Further efforts are needed to test this approach and refine it for sensitivity to observation type, network density, error characteristics, and other factors.

21.6 LONGER TIMESCALES AND SEAMLESS PREDICTION

Numerical prediction beyond the medium range requires coupled atmosphere-land-ocean modelling to account for the more slowly varying processes associated with land surface processes, ocean circulation, and sea ice evolution. Coupling may benefit the shorter ranges as well. In 2008, ECMWF introduced operational coupled ensemble prediction starting at day 11 in its variable resolution EPS system (VAREPS), and in 2014 introduced coupling starting at day 1, representing truly seamless prediction across time scales from the short- to sub-seasonal range. Other major NWP centres are expected to follow suit in due course.

Evaluation of seamless numerical prediction requires verification approaches that allow for consistent interpretation across time scales. This is tricky because short- and medium- range forecasts tend to be deterministic or ensemble predictions of instantaneous[b] "absolute" weather variables at fine spatial and temporal scales, whereas extended range forecasts are based on coarser resolution ensembles, are typically given as probabilistic predictions of weekly or fortnightly anomalies being in a particular category (e.g. highest tercile), and rely on large hindcast datasets for forecast calibration. The variables of greatest interest in the extended range include surface precipitation and temperature, features such as tropical storms and monsoon onset, and indices for modes of variability such as the Madden-Julian Oscillation (MJO).

The verification approach should reflect the way the forecasts are used. In research mode verification of extended range forecasts is generally done against independent observations from surface networks or satellite, or against the hindcast dataset using cross-validation, using standard ensemble and probabilistic diagnostics and metrics like spread-skill plots, reliability and ROC diagrams, and Brier skill score. Real-time verification may compute these metrics for the most recent set of *N* (e.g. 30) forecasts. A challenge for real-time long range forecast verification is estimating robust statistics when the number of forecasts issued in a month or season is relatively small - much smaller than for NWP. For reporting forecast quality to users, simple verification approaches such as percent correct for forecasts above/below the median are often used, but this is not sufficient for model development and improvement.

Seamless verification methods to evaluate medium and extended range models in a consistent way are in their infancy and much more research is needed. A few proposed approaches are mentioned below.

Since the coupled model starts with a set of initial conditions and integrates forward in time, it predicts weather en route to predicting climate. Therefore, verification approaches that are appropriate for weather forecasting can be applied to the shorter-range predictions from coupled models to assess the ability of the model to correctly represent processes. The Transpose-Atmospheric Model Intercomparison Project (AMIP) strategy of verifying climate models in NWP mode is an efficient way to detect errors in model processes that become apparent as biases early in the forecast period (Williams et al. 2013). Modelling centres should include this powerful approach in their programme of model evaluation activities, using as reference data not only standard meteorological observations but also satellite radiances, surface flux measurements, sea surface temperatures, and other non-standard observations.

The real-time multivariate MJO index (RMM) phase plot (Gottschalck et al. 2010) is a climate-focused verification approach that can also be applied to medium-range NWP. There is a need for additional metrics to diagnose other modes of sub-seasonal climate variability in NWP and coupled models.

A seamless approach for comparing forecasts from an extended range prediction system across time scales was proposed by Zhu et al. (2014). They verified 1 day ahead forecasts of 1 day rain accumulation, 2 day ahead forecasts of 2 day accumulation, and so on, out to 4 week ahead forecasts of 4 week rain accumulation. They computed the temporal correlation of observed and

[b] *Rainfall is an exception; short-range quantitative precipitation forecasts (QPFs) are typically accumulated over scales of 1 or more hours*

forecast ensemble mean rainfall at each grid box and found little change in the results whether they used total rainfall or rainfall anomalies. This approach of equivalent lead and aggregation time would also be amenable to verification metrics for categorical, probabilistic, and ensemble forecasts. Depending on the chosen metric (and the verification question it addresses), one could determine the temporal scales with useful prediction skill according to that metric - this would be a promising avenue to explore.

The generalized discrimination score (GDS) described by Mason and Weigel (2009) provides a consistent verification approach across different types of forecasts. Also known as the two-alternative forced choice (2AFC) approach, this method quantifies how well the forecast correctly discriminates between the observations. It has the same meaning when applied to forecasts that are formulated as binary, multi-category, continuous, or probabilistic variables, which can be verified against observations that may be (any of) binary, multi-category, or continuous. The GDS would therefore enable model performance for deterministic short-range forecasts and probabilistic sub-seasonal forecasts of anomalies to be compared in a consistent manner.

Weather represents the rapidly varying flow within a larger scale (climate) regime. Verification of extended range predictions conditional on the climate regime has led to identification of periods of enhanced predictability associated with planetary-scale teleconnections. For example, the MJO phase of tropical convection in the initial state impacts the Northern Hemisphere conditions three weeks later (Vitart and Molteni, 2010). These "windows of opportunity" for enhanced prediction skill are not yet well understood, and require further conditional verification to quantify their benefit in predicting overall weather conditions for applications such as agriculture and water resources.

New applications for extended range prediction will require appropriate verification approaches to be developed. Some examples include windiness / storminess for renewable energy estimation, wave regimes for beach erosion and public safety, heat and humidity conditions for tourism, and so on. New user-relevant metrics will need to be developed in many cases, in close consultation with the relevant sectors.

Many thresholds may be necessary to satisfy a large range of users. For this reason, and for diagnosing and correcting errors in ensemble predictions, verification of the full distribution may be more desirable than simple metrics based on forecasts for terciles or above/below median. The probability integral transform (PIT) and rank histograms can be used to assess the calibration of probabilistic and ensemble forecasts, but research is needed on the best way to evaluate forecasts when the tails of the distribution may be more "valuable" and important to predict correctly. As noted in Section 21.3, the research on methods for verifying probabilistic forecasts of rare extreme events is still in its infancy. Advances in this area are needed to support evaluation of forecasts for extremes across all time scales.

New methods and simple metrics are needed to assess model performance in simulating the climate modes and teleconnections that enhance sub-seasonal predictability. The RMM index for verifying MJO is already in wide use, but other features that require the development and testing of verification metrics include blocking highs, land surface conditions, sea ice concentration and extent, monsoon phase, and storm track variations. Many of these features are coherent structures and may be amenable to the use of spatial verification approaches described in Section 21.2, possibly extended or modified to include the time dimension.

To focus on some of these issues, the community can make use of knowledge gained in the Climate and Ocean: Variability, Predictability and Change (CLIVAR) project and the MJO working group who have focused on the connections between larger-scale phenomena and the performance of forecasting systems. Two of the legacy projects from the WMO's THORPEX (THe Observing System Research and Predictability EXperiment) programme - namely, the Sub-seasonal to Seasonal Prediction Project (S2S, see Chapter 20) and the Polar Prediction Project (PPP, see Chapter 19) - will conduct and apply research on relevant verification methodologies and observations, which should lead to advances in the verification methodologies available for longer-range predictions. The WGNE/WGCM Climate Metrics Panel is developing and promoting a metrics toolkit for verifying the output of coupled climate models, starting with basic quantities like bias and RMSE of key state variables. Diagnostic and process-oriented verification techniques will be included in future releases; many of these techniques may be relevant for assessing extended range predictions.

Improved verification methodologies for extended range forecasts must be tested on large datasets comprising the output of seamless modelling systems, hindcast datasets for calibration and cross-validation, and well-characterized high quality global observations of precipitation, temperature, and other relevant variables. The Obs4MIPS (Observations for Model Intercomparisons) activity, which is making observational products more accessible for climate model intercomparisons, is a good source of non-real-time data (Teixeira et al. 2014). Routine verification in near-real-time (relatively speaking) should leverage current operational seasonal and NWP verification systems.

To advance these developments and applications, verification researchers will need to work with both the short-range and long-range modelling communities who are converging on extended range prediction but still tend to view the world somewhat differently. Verification systems must accommodate different data formats (e.g. GRIdded Binary (GRIB) vs. network Common Data Form (NetCDF)) and temporal/spatial aggregations. Verification of seamless modelling systems must also provide objective evidence to inform the choice that many major centres have between frequent model upgrades to incorporate improvements and boost short-range accuracy, versus freezing a model to accommodate the time consuming generation of hindcasts necessary for calibration.

21.7 ENVIRONMENTAL VARIABLES AND DOWNSTREAM PRODUCTS AND IMPACTS

Seamless prediction also refers to the coupling of weather predictions to other environmental variables such as atmospheric composition and aerosols, streamflow, water quality, and vegetation state. For many years the coupling was one way, but NWP systems such as ECMWF's Integrated Forecast System (IFS) now have the ability to carry some environmental variables directly within the model. Understanding the interfaces and identifying how error sources are propagated from one system to another is critical if predictions are to be improved. Although verification of environmental variables typically uses many of the same statistical metrics and approaches as used to verify meteorological variables, it may be preferable to develop new methodologies targeted to the problem at hand. For example, Demargne et al. (2009) describe diagnostic metrics for verifying deterministic and ensemble hydrologic forecasts that are meaningful for users in the water community.

Weather forecasts inform decision-making in a number of spheres (emergency management, energy, aviation, agriculture, tourism, and many more). A focus area for WWRP is the coupling of weather predictions to downstream impacts. Some centres are doing this automatically by using direct or post-processed NWP model output as input to impact models. An example is the Flood Forecasting Centre in the UK where model output from the variable resolution (UKV) meteorological model is fed directly to the hydrological model for predicting streamflow (Pilling et al. 2014, Lewis et al. 2014). Other examples of downstream impact models include fire spread models and renewable energy generation models.

Impact forecasting raises some interesting challenges for verification. Many of the same issues that arise with verifying warnings of extreme weather (timing, intensity) also apply to warnings of impacts associated with extreme weather. Observations of the impacts may be difficult to obtain for a variety of reasons relating to how they are collected, and by whom, how they are stored and disseminated, and whether they measure something that can be predicted and verified or are only indirectly related to the impact. In some cases that data is purposefully unavailable for commercial or national security reasons. It will be necessary in many cases to strengthen or form relationships with the organizations holding the relevant impact data in order to understand and obtain the data. The opportunities for partnering of meteorological and other agencies can lead to more effective services for the public and other stakeholders.

Communication between the meteorological and various downstream communities is often challenging, with each sector "speaking their own language" and having their own priorities for what makes a forecast useful to them. To enable the benefits of improved weather forecasting to be translated into improvements in downstream impact forecasts, it is necessary to develop and apply verification metrics that are meaningful to the downstream users. The best way to do this is through direct engagement with the users to produce "user-relevant" metrics.

The aviation industry is a heavy user of meteorological forecasts. An example of a jointly developed aviation-oriented verification metric is the flight time error, a measure of forecast upper-air wind accuracy that computes the difference between the observed flight time and the forecast flight time calculated by replacing the actual winds along the flight track with the forecast winds (Rickard et al. 2001). Other examples of user-relevant approaches include the application of spatial techniques to track the occurrence of low-pressure systems, the development of measures to evaluate wind "ramps" for the renewable energy industry, and the measurement of forecast consistency for tropical cyclone track forecasts (Hodges, 1999; Bossavy et al. 2013; Fowler et al. 2015). Methods are also needed to translate verification information into socioeconomic benefits through the application of appropriate cost-loss and other models.

Crowdsourcing and data mining of mobile phone networks, Twitter, Facebook, and other social media, are emerging and promising sources of information that may be used to infer the occurrence, coverage, and impacts of hazardous weather (Hyvärinen and Saltikoff, 2010, Muller et al. 2015). The use of these data for verification is only beginning to be explored. A 2013 study comparing crowdsourced hail observations to official hail reports and severe warning polygons in the United States suggested that, due to biases and inaccuracies related to population density and observer engagement, these crowdsourced data should only be used in conjunction with other databases in order to ensure quality (Pehoski, 2013).

The propagation of errors from the meteorological forecast into the downstream impact forecast needs to be quantified and understood. This requires sensitivity testing, including of the assumptions made by the impact model (i.e. understanding its output errors given perfect input). The longer the chain of models, the more opportunity there is to compound errors. The area that has received by far the greatest attention is the propagation of uncertainty from NWP into hydrological prediction (e.g. Zappa et al. 2010), where it has been shown that the major source of error in hydrological predictions is due to uncertainties in the predicted rainfall. Similar work is required for other hazards to allow greater understanding of the relationships in performance along the forecasting chain.

The application of meteorological verification in additional - but often related - fields also represents a challenge. Just as weather forecast verification methodologies are more advanced than the techniques applied in climate forecast evaluation, these methods are also relevant for other physical and social phenomena for which verification has not traditionally been a key activity. For example, weather forecast verification techniques are being adapted for application in areas such as ocean current and earthquake prediction. Extending our efforts into these areas will undoubtedly lead to new challenges related to users, observations, and methodologies.

21.8 LINKAGES TO OTHER ACTIVITIES

Verification is a critical component of any prediction research and application, and applies to all variables, space and time scales that can be predicted and observed. Each project of the WWRP has a strong verification component that includes both traditional approaches and new improved verification methods that may address some of the more challenging questions about forecast performance. Although the JWGFVR is a focal point for verification research and supports the WWRP in this aspect, there is great interest and innovation in verification throughout the weather and climate community. Some verification research linkages within WWRP and with other activities are noted below.

The Polar Prediction Project (PPP) aims to improve operational forecasting at high latitudes. The baseline forecast performance must first be established through increased attention to verification in the Arctic and Antarctic regions. There are significant observational challenges associated with the sparseness and quality of standard observations, especially in extreme conditions. It is likely that satellites will provide a very important source of data for verification, with model evaluation in "observations space" (simulated satellite radiances) likely to be more robust than using retrieved atmospheric variables for verification. PPP has identified sea ice prediction as one of its priorities, so methods for verifying ice extent and concentration will need to be tested and improved, for both large (basin) and very fine (seaport) scales. Spatial verification methods may prove to be quite beneficial for application to these kinds of predictions. In addition, development of user-relevant methods to verify key polar weather and climate phenomena (e.g. blizzards and fog/visibility) is an important need for this project.

The verification sub-project of the Sub-seasonal to Seasonal Prediction Project (S2S) will recommend verification metrics and datasets for assessing forecast quality of S2S forecasts, and provide guidance for a potential centralized verification effort for comparing forecast quality of different S2S forecast systems, including the comparison of multi-model and individual ensemble systems. Verification research is required to develop user-relevant metrics for sub-seasonal to seasonal forecasts and downstream applications, and to determine how to cope with short hindcast periods and the reduced numbers of ensemble members in hindcasts (compared to real-time forecasts) when constructing probabilistic skill measures. Spatial verification of coherent structures in the anomaly fields will also be explored.

The High-Impact Weather (HIWeather, see Chapter 24) project focuses on weather hazards and their impacts, including how improved high-impact weather predictions can lead to enhanced community resilience through better understanding of risk and vulnerability and more effective communication. Verification research in this project will develop methodologies for verifying predictions of hazard impacts (e.g. floods, transport delays, damage to property, injuries, etc.), and explore the use of new types of observations for verification. The potential utility of crowdsourced and third-party data from social media, sensor networks, and other new technologies is of great interest. With help from the Societal and Economic Research and Applications (SERA) working group, the translation of accuracy improvements into socioeconomic benefit (for example, increased forecast lead time leading to greater opportunities to protect assets and thereby reduce losses) will be investigated for different sectors. Verification methods for nowcasts and high resolution NWP ensembles, especially for predictions of extreme values, will also be tested in this project.

Urban meteorology is receiving increasing attention as numerical models and understanding of urban micrometeorology and atmospheric chemistry enable more accurate very fine scale prediction of weather and environmental conditions. The Global Atmosphere Watch (GAW) Urban Research Meteorology and Environment (GURME) project helps national meteorological centres deal with urban issues, especially air pollution. Traditional verification of forecast concentrations of chemical species and particulates against observations from monitoring sites could be augmented by spatial verification approaches using satellite aerosol optical depth measurements, application of new methods to verify extreme values, and development of user-oriented verification metrics for public health applications. Obtaining adequate observations in the complex urban environment poses a significant challenge for advancing urban meteorology (Carmichael et al. 2014).

Forecast and Research Demonstration Projects (FDPs and RDPs) are useful testbeds for new verification techniques, providing opportunities for verification researchers, nowcast developers, modellers, forecasters, and downstream users to interact closely to improve their respective capabilities. They have provided valuable insights on the utility of real-time verification, and further efforts in this direction should be strongly encouraged.

WGNE has a long history of promoting verification research to help the major numerical modelling centres with model evaluation and intercomparison, and to support WGNE experimentation (recent examples include the Transpose-AMIP, Grey Zone and Analysis Verification projects; WGNE, 2014). WGNE's interest in verification methodologies will remain strong as the capability of numerical modelling systems and data assimilation systems evolve, and new observations become available for data assimilation and verification. The THORPEX Interactive Grand Global Ensemble (TIGGE) dataset will continue to be an important resource for exploring new approaches for verifying ensembles and products derived from ensembles (e.g. tropical cyclone strike probability, heavy rainfall probability, storm tracks, etc.), and addressing issues related to the use of model analyses in verification.

Finally, training and education through workshops and conferences, tutorials and courses, websites, WMO documents, journal articles and other literature, are vital to promote the science and practice of verification. WMO has supported much good work in this area and it is hoped that this can continue.

21.9 SUMMARY AND PROSPECTS FOR THE FUTURE

In the last decade or so the meteorological community has seen the successful development and application of new and improved forecast verification methods for numerical prediction of the earth system; yet many challenges remain.

- Spatial verification methods are becoming mainstream and are in some cases applied operationally as well as in research settings. New research is needed to understand how well these methods apply in regions of complex terrain, for ensemble forecasts, for variables other than precipitation, and how they can be extended into the temporal domain to address timing errors.
- New scores have been developed that provide better ways to compare the ability of forecasting systems to predict extreme events; more experience with application of these scores will lead to their wider use and also to development of additional approaches for this difficult challenge, particularly in the context of ensemble and probabilistic prediction.
- Although standard approaches for evaluation of ensembles have matured and are generally applied in a consistent way, ensemble prediction is still an evolving science and new verification metrics for ensembles continue to be developed. These require testing and refinement for meteorological and downstream applications.
- The development and application of user-relevant approaches for forecast evaluation, as well as the application of methods for downstream forecasts and impacts have blossomed in the last decade; the breadth of possible applications will require some consideration and prioritization in the community, in consultation with relevant external users.
- Development of verification approaches for longer range and seamless predictions have become increasingly important to support new prediction capabilities.
- Incorporation of information about observation uncertainty into forecast verification methodologies remains one of the greatest challenges for our community.

Efforts in these areas are likely to dominate verification research over the next decade.

21.10 ACKNOWLEDGEMENTS

The authors would like to thank Tom Pagano and Gary Weymouth for valuable feedback on earlier drafts of this chapter, and Phil Gill, Anna Ghelli and Sarah Jones for their very helpful reviews.

REFERENCES

Ben Bouallègue, Z. and S.E. Theis, 2014: Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products. *Meteorological Applications*, 21:922-929.

Bossavy, A., R. Girard and G. Kariniotakis, 2013: Forecasting ramps of wind power production with numerical weather prediction ensembles. *Wind Energy*, 16:51-63.

Bowler, N.E., 2008: Accounting for the effect of observation errors on verification of MOGREPS. *Meteorological Applications*, 15:199-205.

Brooks, H.E., M. Kay and J.A. Hart, 1998: Objective limits on forecasting skill of rare events. *19th Conf. Severe Local Storms, AMS*, 552-555.

Carmichael, G., S. Grimmond and H. Lean, 2014: *Urban scale environmental prediction systems*, This volume.

Casati, B., V. Fortin, and L. Wilson, 2014: *A wavelet-based verification approach to account for the variation in sparseness of gauge observation networks*. World Weather Open Science Conference, Montreal, Canada, 16-21 August 2014.

Christensen, H.M., I.M. Moroz, and T.N. Palmer, 2015: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141:538-549.

Ciach G.J., and W.F. Krajewski, 1999: On the estimation of radar rainfall error variance. *Adv. Water Resources*, 22:585-595.

Davis, C.A., B.G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. *Weather and Forecasting*, 24:1252-1267.

De Haan, L.L., M. Kanamitsu, F. De Sales, and L. Sun, 2014: An evaluation of the seasonal added value of downscaling over the United States using new verification measures. *Theoretical and Applied Climatology*, 1-11, http://dx.doi.org/10.1007/s00704-014-1278-9.

Demargne, J., M. Mullusky, K. Werner, T. Adams, S. Lindsey, N. Schwein, W. Marosi, E. Welles, 2009: Application of forecast verification science to operational river forecasting in the U.S. National Weather Service. *Bulletin of the American Meteorological Society*, 90:779-784.

Dey, S.R.A., G. Leoncini, N.M. Roberts, R.S. Plant, and S. Migliorini, 2014: A spatial view of ensemble spread in convection permitting ensembles. *Monthly Weather Revue*, 142:4091-4107.

Dorninger, M., M.P. Mittermaier, E. Gilleland, E.E. Ebert, B.G. Brown, L.J. Wilson, 2013: MesoVICT: Mesoscale Verification Inter-Comparison over Complex Terrain. NCAR Technical Note NCAR/TN-505+STR, 23 pp.

Duc, L., K. Saito and H. Seko, 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus A*, 65, 18171, http://dx.doi.org/10.3402/tellusa.v65i0.18171.

Ebert, E.E. and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, 239:179-202.

Ferro, C.A.T., 2014: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917-1923.

Ferro C.A.T. and D.B. Stephenson, 2011: Extremal Dependence Indices: improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26:699-713.

Fowler, T.L., B.G. Brown, J. Halley Gotway and P. Kucera, 2015: Is change good? Measuring the quality of updating forecasts. *Mausam*, in press.

Gallus, W.A., Jr., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Weather and Forecasting*, 25:144-158.

Gilleland, E., 2010: *Confidence Intervals for Forecast Verification*. NCAR Technical Note NCAR/TN-479+STR, DOI: 10.5065/D6WD3XJM.

Gilleland, E., D. Ahijevych, B.G. Brown, and E.E. Ebert, 2010: Verifying forecasts spatially. *Bulletin of the American Meteorological Society*, 91:1365-1373.

Gneiting, T. and R. Ranjan, 2011: Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business and Economic Statistics*, 29:411-422.

Gorgas, T. and M. Dorninger, 2012: Quantifying verification uncertainty by reference data variation. *Meteorol. Z.*, 21:259-277.

Gottschalck, J., M. Wheeler and co-authors, 2010: A framework for assessing operational Madden-Julian oscillation forecasts: A CLIVAR MJO Working Group project. *Bulletin of the American Meteorological Society*, 91:1247-1258.

Haiden, T., L. Magnusson and D. Richardson, 2014: Statistical evaluation of ECMWF extreme wind forecasts. *ECMWF Newsletter,* No. 139, 29-33.

Hitchens, N.M., H.E. Brooks and M.P. Kay, 2013: Objective limits on forecasting skill of rare events. *Weather and Forecasting*, 28:525-534.

Hodges, K.I., 1999: Adaptive constraints for feature tracking. *Monthly Weather Revue*, 127:1362-1373.

Hopson, T.M., 2014: Assessing the ensemble spread-error relationship. *Monthly Weather Revue*, 142:1125-1142.

Hyvärinen, O. and E. Saltikoff, 2010: Social Media as a Source of Meteorological Observations. *Monthly Weather Revue,* 138:3175-3184.

Johnson, A., X. Wang, F. Kong, F. and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Monthly Weather Revue*, 141:3413-3425.

Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Weather and Forecasting*, 22:637-650.

Lewis, H., M. Mittermaier, K. Mylne, K. Norman, A. Scaife, R. Neal, C. Pierce, D. Harrison, S. Jewell, M. Kendon, R. Saunders, G. Brunet, B. Golding, M. Kitchen, P. Davies and C. Pilling, 2014: From months to minutes-exploring the value of high resolution rainfall observation and prediction during the UK winter storms of 2013/14. *Meteorological Applications*, 22:90-104.

Magnusson L., T. Haiden and D. Richardson, 2014: Verification of extreme weather events: Discrete predictands. *ECMWF Technical Memorandum* 731.

Mason, S.J. and A.P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Monthly Weather Revue*, 137:331-349.

Mass, C.F., D. Ovens, K. Westrick and B.A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, 83:407-430.

Mittermaier, M.P., 2014a: A strategy for verifying near-convection-resolving model forecasts at observing sites. *Weather and Forecasting*, 29:185-204.

Mittermaier, M.P., 2014b: *How temporally representative are synoptic observations*? World Weather Open Science Conference, Montreal, Canada, 16-21 August 2014.

Muller, C.L., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, A. and R.R. Leigh, 2015: Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, doi: 10.1002/joc.4210.

Neal, R.A., P. Boyle, N. Grahame, K. Mylne and M. Sharpe, 2014: Ensemble based first guess support towards a risk-based severe weather warning service. *Meteorological Applications*, 21:563-577.

Pehoski, J.R., 2013: A crowdsourced hail dataset: Potential, biases, and inaccuracies. MS Thesis, University of Wisconsin-Milwaukee, 62 pp. Available at: http://dc.uwm.edu/cgi/viewcontent.cgi?article=1306&context=etd.

Pilling, C., D. Price, A. Wynn, A. Lane, S.J. Cole, R.J. Moore, and T. Aldridge, 2014: From drought to floods in 2012: operations and early warning services in the UK. In: Daniell, T.M, (ed.) *Hydrology in a changing world: environmental and human dimensions.* Wallingford, UK, Int. Assn. Hydrological Sciences, 419-424. (IAHS Publication, 363).

Prates, F. and R. Buizza, 2011: PRET, the Probability of RETurn: a new probabilistic product based on generalized extreme-value theory. *Quarterly Journal of the Royal Meteorological Society*, 13: 521-537.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126:649-667.

Rickard, G.J., R.W. Lunnon and J. Tenenbaum, 2001: The Met Office upper air winds: Prediction and verification in the context of commercial aviation data. *Meteorological Applications*, 8:351-360.

Roberts, N.M. and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Revue*, 136:78-97.

Roulston, M.S. and L.A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Monthly Weather Revue*, 130:1653-1660.

Santos, C. and A. Ghelli, 2011: Observational probability method to assess ensemble precipitation forecast, *Quarterly Journal of the Royal Meteorological Society*, 138:209-221.

Steinacker, R., C. Häberli and W. Pöttschacher, 2000: A transparent method for the analysis and quality evaluation of irregularly distributed and noisy observational data. *Monthly Weather Revue*, 128:2303-2316.

Teixeira, J., D. Waliser, R. Ferraro, P. Gleckler, T. Lee and G. Potter, 2014: Satellite Observations for CMIP5: The Genesis of Obs4MIPs. *Bulletin of the American Meteorological Society*, 95:1329-1334.

Tödter, J. and B. Ahrens, 2012: Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition. *Monthly Weather Revue*, 140:2005-2017.

Tollerud, E., T. Jensen, K. Holub and J. Halley Gotway, 2014: *Points and pixels, gage and sky: Choosing the right observations to verify forecasts.* World Weather Open Science Conference, Montreal, Canada, 16-21 August 2014.

Vitart, F. and F. Molteni, 2010: Simulation of the Madden–Julian Oscillation and its teleconnections in the ECMWF forecast system. *Quarterly Journal of the Royal Meteorological Society*, 136:842-855.

Weijs, S.V., R. van Nooijen and N. van de Giesen, 2010: Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Revue*, 138:3387-3399.

WGNE (Working Group on Numerical Experimentation), 2014: 29th session of the WWRP/WCRP Working Group on Numerical Experimentation (WGNE-29). Melbourne, Australia, 10-13 March 2014.

Williams, K.D., A. Bodas-Salcedo, M. Déqué, S. Fermepin, B. Medeiros, M. Watanabe, C. Jakob, S.A. Klein, C.A. Senior and D.L. Williamson, 2013: The Transpose-AMIP II Experiment and its application to the understanding of Southern Ocean cloud biases in climate models. *Journal of Climate*, 26:3258-3274.

Wilson, L.J. and A. Giles, 2013: A new index for the verification of accuracy and timeliness of weather warnings. *Meteorological Applications*, 20:206-216.

Yamaguchi, M., S. Lang, M. Leutbecher, M. Rodwell, G. Radnoti, and N. Bormann, 2014: *Observation-based ensemble spread-error relationship.* World Weather Open Science Conference, Montreal, Canada, 16-21 August 2014.

Yilmaz, M.T. and W.T. Crow, 2014: Evaluation of assumptions in soil moisture triple collocation analysis. *Journal of Hydrometeorology*, 15:1293-1302.

Zappa, M., K.J. Beven, M. Bruen, A.S. Cofiño, K. Kok, E. Martin, P. Nurmi, B. Orfila, E. Roulin, K. Schröter, A. Seed, J. Szturc, B. Vehviläinen, U. Germann, and A. Rossa, 2010: Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmospheric Science Letters*, 11:83-91.

Zhu, H., M.C. Wheeler, A.H Sobel and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Monthly Weather Revue*, 142:1556-1569.