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

Elizabeth Ebert (Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Australia; e.ebert@bom.gov.au)

Barbara Brown (National Center for Atmospheric Research, Boulder, USA; bgb@ucar.edu)

INTRODUCTION

Numerical weather model forecasts have been verified since the 1950s when they first started providing reasonable predictions. WMO Centers for Deterministic Forecast Verification (located at ECMWF), Ensemble Forecast Verification (JMA), and Long Range Forecast Verification (BOM and MSC) coordinate the routine production of verification results for numerical weather prediction (NWP) models from major national centers. In the case of NWP, the bias, RMS errors, S1 skill score, and anomaly correlation of forecast fields on selected pressure levels have been computed for many decades, and the CBS exchange of standard verification scores for NWP has changed little since it was initiated in 1985. Verification information is also a key ingredient to development of post-processing algorithms for NWP forecasts.

However, NWP models have improved enormously since they were first introduced. Operational global models now have spatial resolutions below 20 km, and it is common for regional NWP to run at spatial resolutions of only a few km. Ensemble prediction is overtaking deterministic NWP as the most important source of numerical guidance, and it too is being run at very high resolution. Coupling of atmospheric models to ocean and land surface models is being brought forward from the seasonal range to shorter ranges, with operational fully coupled NWP soon to become a reality at ECMWF. Numerical weather and climate predictions are also being used to drive downstream impact models for emergency management, hydrology, agriculture, energy, and many other applications.

Improvements in numerical prediction have called for improved methods to verify these forecasts. This has been an active area of research in the last decade or two. The WWRP/WGNE Joint Working Group on Forecast Verification Research (JWGFVR) was established in 2003 to promote work in this area. It coordinates workshops, tutorials, verification methods intercomparisons, and is the focal point for verification of WWRP Forecast and Research Demonstration Projects.

This short paper focuses on some of the recent successes as well as current challenges facing the verification community, as reflected in recent workshops and other presentations and papers.

SUCCESSES, ISSUES, AND CHALLENGES

Spatial methods

Short and medium-range NWP has been evolving toward ever higher resolution and greater emphasis on the prediction of surface weather. The spatial variability and intensity distributions of NWP increasingly resemble observations, including an improving ability to simulate the extreme values that are so important in a forecast and warning context. Although it is tempting to use model output at face value, it is not accurate enough to do that. Traditional verification against standard observations may even suggest that higher resolution NWP is less accurate than lower resolution NWP (e.g., Mass *et al.* 2002). The spatial and temporal scales of the verification have a strong influence on the measured performance, with finer resolution verification 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) describes these methods as neighborhood (crediting "closeness" in space, time, and/or intensity, often through probabilistic approaches), scale separation (quantifying error by scale), 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). An intercomparison of more than a dozen spatial verification methods explored their respective ability to measure location, intensity, and structure errors, distinguish which scales have skill, and verify the predicted occurrence of events. Table 1 gives a summary by type. While all methods measure intensity bias, no single method addresses all types of errors and so 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 errors.

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	×	✓	✓	×	×

Neighborhood and feature-based methods are becoming mature enough to be used routinely in many NWP centers to verify high resolution NWP. For example, the Met Office uses the Fractions Skill Score (Roberts and Lean 2008) and neighborhood Brier score (Mittermaier 2014) to measure the scales at which the model shows useful skill for predicting rainfall and cloud. The Method for Object-based Diagnostic Evaluation (MODE) and the Contiguous Rain Area (CRA) method are used to characterize NWP rainfall performance in the US and Australia (Ebert and McBride 2000, Davis *et al.* 2009, http://www.hpc.ncep.noaa.gov/verification/mode/mode.php#page=page-1). These methods are potentially applicable to other fields of interest (wind, moisture, cloud), and to evaluating timing errors, but more work is needed to explore these possibilities.

Advanced verification methodologies have tended to focus 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 called MesoVICT (Mesoscale Verification In Complex Terrain; Dorninger *et al.* 2013) will test verification methods on precipitation and wind forecasts from deterministic and ensemble NWP, and for the first time include 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.

Extremes

A strong motivation for high resolution NWP is to predict extreme values associated with dangerous weather. One challenge in verifying predictions for extremes is observing them and collecting enough forecast-observation pairs to compute meaningful and robust statistics. In cases of extreme weather the observations may themselves be less trustworthy, for example, when instruments are destroyed by flood or wind or measurements are compromised by the weather conditions.

Some common categorical verification metrics behave badly for rare events, making them ineffective for distinguishing forecast performance. As discussed by Ferro and Stephenson (2011), for imperfect forecasts the threat score and the Gilbert, Heidke, and Peirce skill scores all asymptote to zero for rare events. They propose a new class of scores called extremal dependence scores that reward hits and penalize misses and false alarms as desired, and also behave much more consistently with the forecast performance measured for less rare events. The simplest is the extremal dependence index, $\text{EDI} = (\ln F - \ln H) / (\ln F + \ln H)$, where F is the false alarm rate and H is the hit rate. 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.

For continuous variables, extreme value theory, long used for analyzing extremes in the climate context, may offer some promise (Prates and Buizza 2011) for evaluation of extremes for NWP forecasts. Threshold weighted 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 NWP verification which has fixed base times and valid times, warning verification has additional requirements to evaluate lead time and warning duration relative to the onset and cessation of the event being warned for. The trade-off between lead time and warning accuracy needs to be assessed, which can inform user-focused studies of warning effectiveness in the face of false alarms. The spatial extent of the 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 neighborhood 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 better understand the warning performance as a function of scale and other factors. 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).

Ensembles

Ensemble prediction is now being used at all scales to explicitly account for forecast uncertainty related to initial conditions and model uncertainties. Traditional verification of ensemble prediction systems (EPS) uses metrics such as the Brier skill score and continuous ranked probability score, and diagnostics such as reliability and relative operating characteristic diagrams and rank histograms, to assess spread-error consistency and reliability and discrimination of probability forecasts. 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.

The Lead Center for Verification of EPS at JMA collects reliability tables from national centers running EPS and produces a suite of verification statistics following guidelines from the *Manual on the GDPFS (WMO-No.485)*. Not all centers take advantage of this service, but as ensemble predictions are used for a wider variety of meteorological and downstream applications it is more important than ever to know their accuracy.

Spatial methods are now starting to be used to evaluate ensemble predictions. Neighborhood methods are easily extended to include an ensemble dimension (e.g., Duc *et al.* 2013). Approaches for feature-based ensemble verification are still being investigated. 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.

Uncertainty

Uncertainty in verification arises from many sources. Perhaps most importantly, (a) observations are inherently uncertain due to measurement and representativeness errors, and (b) forecast verification is applied to samples of forecasts, which leads to uncertainty related to sampling variability. Sampling variability is somewhat more straightforward than observation-related uncertainty to cope with, 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, or 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 error, some knowledge has been gained in recent years. However, much more information is required. Fundamentally, as models 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 dealing with observation error. 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.

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). 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.

Longer time scales and seamless prediction

Numerical prediction beyond the medium range requires coupled atmosphere- ocean/ice modelling to account for the more slowly varying processes associated with the ocean circulation. Coupling may benefit the shorter ranges as well. In 2008 ECMWF introduced operational coupled ensemble prediction starting at day 11 in its variable resolution VAREPS system, and will soon have full coupling starting at day 1, representing truly seamless prediction across time scales from the short- to sub-seasonal range. Other major NWP centers 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[1] "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, Brier skill score, etc. Real-time verification may compute these metrics for the most recent set of N (e.g., 30) forecasts. 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 modelling in a consistent way still

[1] Rainfall is an exception; short-range QPFs are typically accumulated over scales of 1 or more hours.

need further research. A few 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-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). 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 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 recently proposed by Zhu *et al.* (2013). 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 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.

The generalized discrimination score described by Mason and Weigel (2009) provides a consistent verification approach across different types of forecasts. Also known as the two-alternative forced choice, this approach quantifies how well the forecast correctly chooses between any two 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 generalized discrimination score 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 identified periods of enhanced predictability associated with planetary-scale teleconnections. For example, the MJO phase of tropical convection in the initial state impacts the N. 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 verification to quantify their benefit in predicting overall weather conditions for applications such as agriculture and water resources.

When verifying forecasts for extreme conditions, which are normally given as probabilistic forecasts for upper quantiles (e.g., top 5%), adequate sample sizes may be difficult to obtain and it is necessary to compute confidence intervals on the verification results. This information could assist decision makers in making best use of extended range forecasts for extreme conditions where impacts are likely to be greatest.

Applications to 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 IFS now have the ability to carry some environmental variables directly within the model. Verification of environmental variables typically uses many of the same statistical metrics and approaches as used to verify meteorological variables. Demargne *et al.* (2009) describe diagnostic metrics for verifying deterministic and ensemble hydrologic forecasts that are meaningful for users.

Weather forecasts inform decision making in a number of spheres (emergency management, energy, aviation, agriculture, tourism, and many more). A focus area for WWRP and a key topic in WWOSC2014 is the coupling of weather predictions to downstream impacts. Some centers 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 UKV meteorological model is fed directly to the hydrological model for predicting streamflow. 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.

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 (which is not truly a verification metric since it doesn’t involve the observation; yet it is an important factor for forecasters and end users of the predictions).

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.

SUMMARY AND PROSPECTS FOR THE FUTURE

As discussed in this paper, the development and application of forecast verification methods for NWP predictions has experienced many successes in the last decade; yet many challenges remain. For example, 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, and for variables other than precipitation. 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. Standard approaches for evaluation of ensembles have matured and are generally applied in a consistent way. 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. Incorporation of information about observation uncertainty into forecast verification methodologies remains one of the greatest challenges for our community, along with development of verification approaches for longer range and seamless predictions. Efforts in these areas likely will occupy verification efforts over the next decade.

REFERENCES

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.

Ciach G.J., and W.F. Krajewski, 1999. On the estimation of radar rainfall error variance. *Advances in 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. *Wea. Forecasting*, **24**, 1252-1267.

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. *Bull. Amer. Meteorol. Soc.*, **90**, 779–784.

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. *J. Hydrology*, **239**, 179-202.

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

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. *Bull. Amer. Meteorol. Soc.*, **91**, 1365-1373.

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

Gottschalck, J., M. Wheeler and co-authors, 2010: A framework for assessing operational Madden–Julian oscillation forecasts: A CLIVAR MJO Working Group project. *Bull. Amer. Met. Soc.*, **91**, 1247-1258.

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

Mason, S.J., and A.P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331-349.

Mass, C.F., D. Ovens, K. Westrick and B.A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Met. Soc.*, **83**, 407-430.

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

Prates, F. and R. Buizza, 2011: PRET, the Probability of RETurn: a new probabilistic product based on generalized extreme-value theory. *Q.J.R. Meteorol. Soc.*, **137**, 521–537.

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. *Meteorol. Appl.,* **8**: 351-360.

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

Rodwell, M.J. and T.N. Palmer, T. N., 2007: Using numerical weather prediction to assess climate models. *Q.J.R. Meteorol. Soc.*, **133**, 129–146.

Vitart, F. and F. Molteni, 2010: Simulation of the Madden–Julian Oscillation and its teleconnections in the ECMWF forecast system. *Q. J. R. Meteorol. Soc.*, **136**, 842-855.

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. *J. Climate*, **26**, 3258-3274.

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. *Atmos. Sci. Lett.*, **11**, 83–91.