
Environmental Prediction Systems: global and medium-range aspects

This article summarizes recent progress in global Environmental Prediction systems, focusing on the development of numerical models, the use of ensemble prediction and the performance of global systems in the medium-range. It provides expected directions for the future.

Florence Rabier, Alan Thorpe,

ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX United Kingdom

Andy Brown (Met-Office), Martin Charron (Environnement Canada), Tom Hamill (NOAA), Junichi Ishida (Japan Meteorological Agency), Bill Lapenta (NCEP)

ABSTRACT

Over the past 30 years the skill of global numerical weather predictions has significantly improved: high-resolution global forecasts now routinely exceed a defined useful level of skill up to $\sim 6\frac{1}{2}$ days ahead, with particular forecasts extending considerably further. The rate of improvement continues at about 1 day per decade of research and development. Weather forecasts involve accurate and reliable ensembles of numerical predictions defining: the most likely future weather, a quantitative measure of the confidence that can be placed on that most likely outcome, and definition of the range of less likely but still plausible scenarios. This white paper considers how this progress has been made and it provides a picture of future scientific opportunities by covering the: importance of resolution, physics, and coupling; use of ensembles/reforecasts; assessment of performance using standard scores and estimation of performance for severe weather.

INTRODUCTION

The advances in global NWP made in the past decades have arisen from scientific developments that have:

- reduced numerical errors through more accurate numerical methods and increased spatial resolution, enabled by increasing supercomputer capacity;

-
- improved the quality of the initial conditions by developing data assimilation methods that combine the increasing number and variety of observations with prior information from forecasts;
 - improved the representation of physical processes, using basic meteorological research on: clouds, convection, sub-grid scale orographic “drag”, surface interactions, etc;
 - enabled the design of reliable ensemble predictions through the inclusion of initial condition and model uncertainties such that probabilities can be inferred.

Numerical weather prediction is now based on the underpinning concept of estimating the initial-time probability density function and predicting its evolution by using an ensemble of realizations of the system. A state of the art global forecasting system in 2014 operates with around 20-50 ensemble members and a horizontal resolution in the range 16 to 50km with of order 100 vertical levels. The initial ensemble spread of Z500 for the northern hemisphere is about 2.5m (or only about 3% of the variability) with an initial exponential growth rate of the spread in the forecast of about 1 day^{-1} . On average by 10 days into a forecast this spread has grown to around 70m.

Looking forward, the prospect for global NWP is that it will further approach being at kilometre resolution so that global forecasts can be closer to the human scale. We can aspire to predict large-scale weather patterns and regime transitions out to a month or more ahead and high-impact events out to two weeks ahead both accurately and reliably. There are good indications that, under certain conditions, global anomalies could exhibit predictable signals on seasonal time-scales.

It is now apparent that many components of the Earth-system (e.g., atmosphere, oceans, composition, cryosphere) are influential for medium-range weather predictions. Also analyses and predictions of these components, on a range of time-scales, have societal significance and so numerical weather prediction is evolving into numerical environmental prediction. Coupling the components of the Earth-system, including the data assimilation, is becoming a major aspect of the future science that is needed.

DEVELOPMENT OF ENVIRONMENTAL PREDICTION SYSTEMS

Background:

Dramatic progress in the performance of global numerical weather predictions is reported elsewhere in this document. As noted in the introduction, improved observational coverage, observation quality and data assimilation algorithms have been an important contributor to these improvements. In addition, developments to the models themselves have also been crucial. These have included both resolution (horizontal and vertical) and improvements to the representation of physical and dynamical processes.

The ability of higher horizontal resolution models to have smaller truncation errors and to better represent processes, weather systems and surface forcing (e.g. better resolved topography) has consistently been found to improve performance. Hence large amounts of increased computer power have been invested in increasing resolution, although of course balanced judgments need to be made about the best use of resources (e.g. trade-offs between the computational costs of resolution, ensemble size, model complexity and data assimilation).

Models have also improved their representation of the vertical structure of the atmosphere, both by raising the tops of the models (often now located between 0.1 and 0.01 hPa, i.e. between about 65 and 80 km altitude) and by increasing the number of vertical layers. One of the key benefits of the former is that more accurate profiles in the stratosphere allow better use of satellite data and hence improve the quality of the analysis (and hence the forecast). There is also some evidence that a better representation of the stratosphere can directly improve tropospheric forecasts (Roff et al, 2011, Shaw and Shepherd, 2008, Tripathi et al, 2014), although assessing the relative importance of this on different timescales is still an area of active research. Increasing vertical resolution in the troposphere has often proved a difficult change for operational centres to successfully make, requiring some retuning of model physics in order to achieve satisfactory performance. In part this may indicate undesirable resolution sensitivities of the physics schemes, but in part it may simply be indicative that the vertical resolution remains insufficient to properly represent important processes and phenomena (e.g. relatively shallow layer clouds).

Improvements to the representation of physical processes have also been important not least because of their influence on the large-scale circulation patterns. For example the accuracy of NWP forecasts is hugely sensitive to the (still relatively uncertain) representation of surface drag, and errors in the representation of convection can have significant remote influences in the medium-range.

A key aspect of the global models that are being used for NWP is that they are increasing in complexity in the sense that there are other components of the Earth-system that are included in addition to the atmosphere. These include the oceans, the land surface, atmospheric composition, sea-ice, etc. This is motivated because research is indicating that these other components contain sources of weather predictability, e.g., long-lived anomalies in soil moisture, sea-surface temperature, and sea-ice. It is also motivated by the fact that society and decision-support agencies require analyses and predictions of aspects of these components, e.g., atmospheric composition for air quality and GHG monitoring, ocean state, and flooding. The growth of Earth-system science and a holistic approach to the natural environment has developed most strongly in the climate science community but is growing rapidly in weather prediction also. It means that many scientific disciplines other than meteorology now are involved in the scientific and modelling developments that are needed. These include: atmospheric chemistry, oceanography, hydrology, glaciology, sea-ice science. The interactions between the weather and the physical, chemical and biological properties of the system can lead to complex inter-connections. For example, there is evidence that the evolution of hurricanes on time-scales of 3-7 days can be significantly influenced by the presence of a coupled ocean in numerical prediction models. It has also been shown that accurate treatment of aerosols in an NWP model can affect wind speeds via the radiative forcing and this in turn can affect weather phenomena such as heavy rainfall within the Indian summer monsoon. Another

potential example is a link between the rapidly changing sea-ice coverage of the Arctic basin which is believed to have a role in affecting the northern hemisphere circulation and so the predictability of European weather. This has led NWP centres to add interactive components such as a coupling to an ocean circulation model even from day zero in weather predictions. This will present increasing and exciting scientific challenges such as devising coupled ocean-atmosphere data assimilation methods and ways to represent the complexity of tropospheric chemistry without prohibitive computational cost.

Underpinning Research:

Looking forwards, the steady progress that has been made over the past decades to reduce horizontal mesh sizes is expected to continue to reap the associated benefits. In order to do this, there are specific science challenges that will need to be addressed. These include how to transfer from parametrised to explicitly resolved processes, such as those associated with deep convection. Also, data assimilation will have to transition to become fully multi-scale/multi-parameter schemes for the various components of the future coupled system.

In order to make progress, significant developments in many aspects of the model representation of physical processes are required. For example, although an ‘old’ problem, representation of the stable boundary layer remains problematic, with challenges to achieve realistic near-surface temperatures while at the same time achieving good synoptic behaviour. In part at least the latter may be related to issues with the drag parametrizations. Current work co-ordinated by WGNE has revealed that while different leading operational centres typically have very similar zonal mean drags (as scores degrade very quickly if this is not optimized), they achieve this through very different combinations of boundary-layer and orographic drag, suggesting a fairly arbitrary tuning of schemes against each other. A real challenge is to try to come up with techniques to better disentangle compensating errors (both in drag and more widely). The use and detailed analysis of errors in short forecasts e.g. day 1 (or even in the limit the first time-step) can certainly help in this process as errors remain more linear and closer to source. However, further assessments of and direct constraints on individual schemes (e.g. from observations or from using high resolution models as surrogate truth) are also required.

The representation of (tropical) convection is another area that remains particularly challenging, with most global models struggling with convective organization and the diurnal cycle (although some progress is being made). Indeed it seems plausible that making significant progress may require challenging some of the traditional paradigms for parametrization (such as treating each column individually), with future schemes likely to have to represent organization across multiple columns, have memory and an in-built representation of uncertainty (Holloway et al, 2013). They will also have to be scale-aware and able to cope with the problem of convection becoming partially resolved (an area that is undergoing active current research such as via the WGNE grey-zone project).

Other areas worthy of increased attention include the numerics of many of the physics

schemes (e.g. microphysics), and the coupling together of the physics and dynamics. Furthermore, many operational models show spectra that tend to fall-off more rapidly than observed at scales a surprisingly long way above the grid scale (e.g., six to eight times the mesh size). The reasons for this are not fully understood, and there are certainly implications for the physical schemes.

The tuning, over the last 20 years, of initial condition and model error uncertainty, has allowed mean ensemble spread to approach mean ensemble error for upper-air parameters. The challenge for the future is to do this on a flow-dependent basis. In order to do this, an important area of research is to utilise our knowledge of the uncertainties in individual physical processes to generate the model uncertainty component of ensemble design. Today there is usually either no link or limited connectivity between the physical parametrisations and the representation of model uncertainty via the variety of schemes used in operational prediction systems that are sometimes referred to as “stochastic physics”. Indeed the whole area of model uncertainty is one where substantial progress is needed if this vital element in generating forecast errors and unreliability is to be properly addressed. Another example of the large effective resolution issue referred to earlier is that current stochastic physics schemes have to use long correlation space scales in order to impact the ensemble spread appropriately.

A continuing and important area of research is regarding the sources of predictability in the Earth-system. To use terminology that Vilhelm Bjerknes would have recognised in 1904 – predicting future weather really is a battleground with the forces of predictability pitched against those of unpredictability. The sources of predictability include: large-scale forcing of smaller-scale weather; teleconnections or the chain of predictability; long-lived coherent structures. The sources of unpredictability include: upscale energy propagation and instabilities injecting chaotic “noise”; errors in numerical and physical approximations; insufficient number and poor use of observations. The outcome of this battleground could be described in terms of noise growing during the forecast and thereby leading to limits to predictability. The conventional wisdom might suggest that the limit is around two weeks ahead. But we need to ask what are the predictable signals and on what time-scales - is there music lurking within that noise? Coherent long-lived phenomena (and propagating Rossby waves) provide predictability and space-time averaging isolates predictable signals. This has been referred to as “predictability in the midst of chaos”. It suggests that the concept of a limit to predictability be replaced by the concept of a seamless predictive capability on a wide variety of time-scales (see Hoskins, 2012). Space-time average properties exhibit much longer predictable time-scales. Prospects over the next decades might be characterised as: global NWP at kilometre horizontal resolution (towards “forecasts on the human scale”) by 2030; accurate and reliable prediction of high-impact weather out to 2 weeks ahead; prediction of large-scale weather patterns and regime transitions out to a month or more ahead; prediction of global circulation anomalies out to a year ahead.

Linkages:

An active physics community is already co-ordinated through GEWEX GASS (and

GLASS for the land surface), which runs numerous projects typically bringing together observations and process models (e.g. cloud resolving models) and using them to try to understand and improve the performance of operational models. Although hosted under WCRP, the fast physics issues are almost entirely common for weather and climate models, and hence this grouping serves the ends of both the weather and climate communities. Many of these projects do successfully involve the academic community, although there remain challenges in truly getting academic actively involved in model development (as distinct from model evaluation or process research). In part these challenges are technical, and need to be overcome through close partnership working between individual centres and academia, but there are wider issues (being considered by the WCRP Modelling Advisory Council) including whether career paths in academia really encourage individuals to get involved in detailed model development.

WGNE is also very active in bringing together modelling centres, sharing progress and running projects to tackle problems of common interest (e.g. current studies include drag comparison; grey-zone; assessment of impact on forecasts of different levels of aerosol complexity). It also provides a vehicle to link to climate expertise that is becoming increasingly valuable to the NWP community (as well as vice versa). For example, as NWP models move towards coupled oceans there is clearly much to be learned from experiences with coupled seasonal and climate models. The new S2S project also plays a valuable role in bridging between weather and climate. Also as the top of NWP models have been raised, further assessment of their performance in the stratosphere – and of how best to represent non-orographic gravity waves – can be best (and is increasingly being) done in close collaboration with the active climate community.

The societal benefits from the Earth-system or environmental approach are substantial. In Europe a major new programme involving several billion Euros of funding is being supported by the European Union - it is the Copernicus programme. This is to provide a single authoritative source of quality-controlled information on the state of the global natural environment for policy-makers and businesses. This is a continental and therefore multi-national collaborative response to a burgeoning environmental information service industry that is developing. A backbone of the Copernicus programme is a European addition to the global observing system - the Sentinel satellites - adding to the other weather and climate measurements made routinely and for special purposes.

Another example of such global environmental prediction systems driven from a societal need is provided by the Canadian Global Ice Ocean Prediction System (GIOPS, see Smith et al., 2014). Marine traffic in the Arctic is increasing significantly, and the demand for atmosphere-ocean-ice forecasts is being amplified by the increased economic activities in this region. GIOPS comprises ocean and ice assimilation systems, and provides 10-day forecasts of ocean-ice conditions at a grid spacing of 0.25°. Currently, a one-way coupling with the atmosphere is operational at the Canadian Meteorological Centre, but a fully coupled atmosphere-ocean-ice system is in development and should become operational within a few years.

Requirements:

One of the biggest challenges in the coming years will be coping with significantly different supercomputer architectures. In the past, developments have been primarily science driven, with algorithm optimization following. However, increasingly the need to consider what will run efficiently needs to be taken into account much more upfront in the choice and design of algorithms, and this will require much closer working between scientists and computational specialists. Many centres are already looking into this in the context of dynamical core design (e.g. choices of grids; choices of implicit versus explicit methods, choices of advection methods), but similar considerations will need to be taken across all aspects of our modeling systems - the term “scalability” has been coined for this variety of critical aspects of the NWP-computer system. Given the scale of the challenges here, almost certainly too large for any one centre to tackle alone, there is a need to consider how international co-ordination can help e.g. through sharing experiences, through jointly developing algorithms and in interacting with vendors.

USE OF ENSEMBLES AND REFORECASTS

Background

Most weather prediction centres now routinely generate regional and/or global ensemble predictions, with multiple realisation or “scenarios” being generated for each forecast to span the uncertainty space. The regional systems are now commonly generated at grid spacings < 10 km and up to 1-2 days, with several using forecast modeling systems employing convection-permitting models and resolutions < 5 km. The regional models are relied upon for providing situational awareness of the likelihood of high-impact weather at the mesoscale, such as heavy rainfall, severe local storms, and tropical cyclone intensity. Global ensemble prediction systems are routinely run into the medium range (< 2 weeks). In 2014, most employ grid spacings of ~ 20 -80 km, with the smallest being 16km. A state-of-the art system will have tens of ensemble members with the highest in 2014 being 52 members. These models are used to provide estimates of the synoptic-scale uncertainty, as well as to provide probabilistic guidance on high-impact events such as tropical cyclones. Both regional and global models have also been coupled to other components of the Earth system, such as to ocean models to predict wave heights and to hydrologic models to provide probabilistic estimates of streamflow.

Most current ensemble prediction systems under-estimate the forecast uncertainty except for the larger-scales; their spread (the standard deviation of the ensemble about its mean) is smaller than the ensemble-mean error, though the two should be consistent in magnitude on average. The two underlying reasons for forecast uncertainty are: (1) the rapid growth of forecast errors from initially small errors due to chaos (Lorenz 1993), and (2) the uncertainty contributed by the use of imperfect, and frequently deterministically formulated prediction systems. The aim for an ensemble forecast is to be both accurate (in the sense that a measure of forecast error is small) and reliable (in the sense that the predicted frequency of occurrence matches the observed frequency of a given event).

Since users expect ensemble guidance to provide useful estimates of the situation-dependent uncertainty, and since there are substantial challenges to designing

ensemble prediction systems to correctly address (1) and (2) above, users have tried some other conceptually simpler methods for achieving more reliable probabilistic predictions. One method is multi-model combination; ensemble prediction data is shared between operational centres, and products are derived from combined guidance (e.g., Fig. 1). This data sharing occurs operationally between the US and Canada through the NAEFS (North American Ensemble Forecast System), and the WMO/THORPEX has promoted experimental product development of multi-model products through its TIGGE (THORPEX Interactive Grand Global Ensemble) project (Bougeault et al. 2009). Global ensemble forecasts are available, with a two-day delay, from web sites at the National Center for Atmospheric Research (NCAR) in the US, from the European Centre for Medium-Range Weather Forecasts (ECMWF) in Britain, and from the China Meteorological Administration. The TIGGE database can conveniently provide data to those scientists wishing to understand the merits of multi-model vs. single-model ensemble prediction systems (e.g., Hagedorn et al. 2012, Hamill 2012). Regional multi-model collaboration has also been facilitated by the WMO through the TIGGE-LAM (Limited-Area Model) project. Though multi-centre ensemble combinations have been demonstrated in many circumstances to improve skill and reliability, they are ensembles of convenience. They have not been explicitly constructed to simulate all the sources of initial-condition and model uncertainty from first principles. Further, their improvement depends on the improvement of the constituent ensemble prediction systems.

Another method for improving on forecast reliability of existing ensemble prediction systems is through statistical post-processing. Past forecasts and associated observations or analyses may be used to determine statistical adjustments to apply to the current forecast. For some fields such as 2-meter temperature at short forecast lead times, a relatively modest sample of a few months provides enough data to substantially improve upon the forecast. For other variables such as heavy precipitation, severe weather, or longer-lead temperature forecasts, one typically notices that the statistical post-processing can be improved by having a much larger training sample. When an assimilation/forecast system is frozen and past forecasts are generated using that frozen system, these are commonly called “reforecasts,” or “hindcasts” in the climate community. Many operational centres have experimented with the generation of reforecasts, including the US National Weather Service, ECMWF, Meteo-France, and the Canadian Meteorological Center. Logistically, reforecasting can present challenges; the computational and personnel expense of reforecasts and associated re-analyses is significant. Freezing the operational model and/or data assimilation system to avoid reforecast re-generation can unacceptably slow the rate of improvement of the raw forecast guidance, but an older reforecast data set that is statistically inconsistent with the new real-time guidance is of little value. The improved skill and reliability from reforecasts, however, is so substantial (e.g., Fig. 2) that many centres are attempting to provide them despite the logistical hurdles.

Underpinning research

Were we able to produce reliable, skillful ensemble guidance, there is bountiful evidence that improved user decisions could be made based on this probabilistic information (e.g., Zhu et al. 2002). There are several hurdles to realizing these improved decisions. First, the ensemble prediction systems do not routinely generate

sufficiently reliable forecast guidance, and this problem is worse for high-impact weather elements such as heavy precipitation than it is for commonly referenced mid-tropospheric elements like 500 hPa geopotential height. Hence, research to improve the ensemble prediction systems such that they properly simulate the uncertainty related to both initial condition errors and the model uncertainty are critical, especially methods that realistically simulate the uncertainty related to these high-impact events. Statistical post-processing methods could be improved as well. In particular, it would be helpful to employ post-processing methods that are efficient with the training data, i.e., that produce reliable guidance even when using only a modest number of past reforecasts.

Many users and even some forecasters are not yet comfortable with making decisions based on probabilistic guidance; they are more comfortable with more definite guidance (“cloudy, with a high of 18C tomorrow”). Education is needed for forecasters and users: what causes forecast uncertainty, how to interpret ensemble guidance, how to make improved decisions based on that guidance. In many cases, too, the ensemble information is yet not synthesized in such a way as to be maximally useful to the forecaster or decision maker. Hence, research and development is also needed into how to present ensemble information in convenient ways to the end user, so as to best facilitate their particular decision-making process.

The US Hazardous Weather Testbed may provide a useful paradigm for making progress on several of these research fronts simultaneously. During the US spring tornado season, several modeling groups contribute high-resolution, storm-resolving forecast guidance to scientists and forecasters gathered at the US Storm Prediction Center (SPC). The scientists and forecasters evaluate ensemble predictions and use them to make experimental forecast guidance of severe-storm potential. They evaluate their prior forecasts, learning how they could have improved on their communication of forecast risk, and they also provide feedback to model developers on the strengths and weaknesses of the modeling systems from the perspective of their ability to simulate the severe weather and its uncertainty. Programs like this thus serve the dual purpose of educating forecasters on how to best leverage the (yet somewhat unreliable) forecast data, while model developers get practical feedback on model performance in aspects that are of greatest societal relevance.

Linkages

The THORPEX legacy “High-Impact Weather” (HIWeather) project is oriented around issues such as have been discussed here, facilitating the use of forecast guidance for making improved decisions related to high-impact weather and to further improving the ensemble guidance that is used in making these predictions. There are other WMO-sponsored programs that address particular forecast challenges. Forecasting problems related to Arctic and Antarctic prediction, for example, are addressed through the collaborative R&D facilitated by the WMO’s Polar Prediction Project and Sub-seasonal to Seasonal Prediction Project. Improving predictions and predictive skill of severe weather forecast capacity in developing nations is addressed through the WMO’s Severe Weather Forecast Demonstration Project.

Requirements

Hirschberg et al. (2011) provides more requirements for further research and development to fully utilize data from ensemble prediction systems. Though oriented around US ensemble prediction deficiencies, the plan provides a convenient outline of the many components needed to realize effective usage of uncertainty information. These include a necessity to better *understand forecast uncertainty*, doing the research to quantify predictability related to high-impact phenomena and to identify the societal needs and best methods for communicating forecast uncertainty information. Another requirement is to *generate improved forecast uncertainty data, products, and services*. To achieve this, we will need to do the research and development to improve the ensemble prediction systems, the post-processing, the verification. Particular attention should be paid to the development of methods that unify the data assimilation and initial condition generation (e.g., ensemble Kalman filters and their hybridization with variational assimilation methods). Developing physically based methods of estimating the model uncertainty are also very important. We should *upgrade the supporting infrastructure*, including improved high-performance computing as well as the storage space needed for the archival of forecast, observational, and analysis data, and the bandwidth to transmit this data to forecasters and users. We'll need to improve the methods for displaying the much more voluminous and complicated ensemble information. Finally, we will need to *communicate forecast uncertainty information effectively*. This will involve reaching out to, informing, educating, and learning from users, training atmospheric scientists in uncertainty quantification and how to communicate this, and developing products that tailor the ensemble-related information to be maximally useful for making improved decisions.

PERFORMANCE OF GLOBAL ENVIRONMENTAL PREDICTION SYSTEMS

Background:

In the last twenty years, progress in Numerical Weather Prediction has been tremendous, as illustrated for example in Figure 3 for the ECMWF model. As mentioned earlier, this progress was driven by improvements in models (resolution and description of physical processes) and in a better use of an increased number of observations. Although it is relatively easy to evaluate these improvements over long time-series, the year-to-year improvement is more subtle and performance depends not only on improvements in the forecasting system but also on the intrinsic predictability and degree of activity of the atmosphere in specific regions. It is thus particularly relevant to compare the actual operational performance at a given time with that of a reference system, which is fixed over a few years. This can be achieved by running forecasts as part of a reanalysis system, and comparing those with the operational forecasts, as shown in Figure 4.

As increasingly forecasters rely heavily on probabilistic forecasts, the performance measures refer to the performance of the ensembles. These are sometimes difficult to interpret as they depend on the improvements in the underlying data assimilation and model and on the enhancements of the ensemble system itself through a better representation of uncertainties. In this situation too, a reference system with respect to which one could compare the performance would be needed. At the moment, as

re-forecasts are not computed with the full-fledged system (same number of ensemble members in particular), one uses instead the TIGGE archive to compare between various centres.

Measuring average performance is important but evaluation of individual forecasts during case studies of extreme events is also crucial. For example, in 2012, Hurricane Sandy hit the eastern coast of the USA with major disruption and casualties. Such cases are the ones for which the quality of the forecast can make the difference in public response and mitigation of weather-related impacts. For this reason, they are investigated in great detail in major NWP centres (Magnusson et al, 2014). Diagnostics obtained on these cases of extreme weather are useful to understand better how the forecasting system behaves “under pressure”, and how the different components fit together. Extreme weather often includes small-scale structures, rapid development and large observation departures. Investigations of the system during extreme situations might highlight deficiencies that also impact normal weather, but are enhanced in these extreme conditions. Case studies can also inform us about ensemble forecast deficiencies, particularly in cases of small spread and large error.

As systems get more coupled/integrated, the overall performance is also increasingly measured through the various components (hydrology, ocean, land, cryosphere, atmospheric composition). For example, flood forecasting can reveal deficiencies in the precipitation amounts produced by the weather forecast, or ocean forecasting deficiencies in surface fluxes.

Underpinning Research:

Evaluation of the forecast performance is generally performed on a few main parameters, verifying against own analyses or observations. This is extremely useful as these standardized scores are exchanged globally and give a broad description of the main operational systems at any given time.

The verification against an analysis has many advantages. Namely, the analysis is an optimal combination of all the available information (observations and model), with generally high accuracy. Furthermore, the analysis is global, at the same resolution as the model forecast. However, the use of its own analysis to verify a forecast has some caveats, as the analysis depends on the forecast model itself, and in particular on its biases. The performance of the forecast at short lead-times can be over-confidently assessed in regions where the analysis is not sufficiently constrained by the observations, such as the stratosphere, the polar or tropical regions. Research is needed on the use of other analyses such as a consensus analysis or analyses randomly drawn from a set of different systems.

Verifying against observations is a very relevant approach. However, the radiosonde coverage is too inhomogeneous to be globally representative and one should consider relying on other observations such as the satellite measurements. In particular, progress is needed on verification with respect to satellite measurements for cloud and radiation evaluation. There is also a need to add to large-scale parameters such as geopotential at 500 hPa or temperature at 850hPa, and each centre verifies a

whole range of weather parameters such as precipitation, surface temperature or wind gusts. As resolution increases, the observations needed on the global scale will need to be much more numerous than the ones currently exchanged internationally, both for verification and data assimilation.

Ensemble verification is also an area which will benefit from research and development to quantify all aspects of reliability (agreement between forecast probability and mean observed frequency), sharpness (ability to forecast probabilities which are not clustered around the mean) and resolution (ability to resolve the set of sample events into subsets with characteristically different outcomes).

For extreme events, as these are rare, many of the scores degenerate to non-meaningful values. In Ferro and Stephenson (2011) the symmetric extremal dependence index (SEDI) was introduced, which does not have this property and research needs to be pursued in this area, together with the investigation of individual cases.

Although the general forecast performance has significantly improved in the last decades, there are still occasional forecast busts. For example, Figure 5 shows anomaly correlation time series of Z500 over Europe at a lead time of 6 days for single forecasts started at 0000 and 1200 UTC by several of the world's weather prediction centres. In general, scores fluctuate about the 80% level, but between 7 and 10 April 2011, there was a strong drop in performance. The frequency of these busts has decreased in parallel with forecast improvement. However, even a low level of busts causes problems for users of NWP products and motivates dedicated investigations in order to understand the potential issues associated with them. Research on how to best perform these investigations is active. One approach used by Rodwell et al (2013) is to make a large composite of several hundred bust events in order to identify common features. Composites of ensemble (rather than single) forecasts also allow us to estimate predictability and to relate this to forecast error.

An area where users would benefit from improvements in the forecast is the prediction of changes in weather regime at the medium to extended range. The forecast performance has to be diagnosed in relation to these weather regimes in order to identify specific deficiencies in the forecasts such as the under-prediction of blocking situations for example. This can raise awareness at the user level and can also be used to diagnose which deficiencies in the model can be responsible for poor performance. Other aspects are the understanding of the processes and interactions between scales, and the identification and diagnostics of the processes that are relevant to forecast performance.

Verification is used to set performance targets and therefore plays a major role in system development. It is essential that scores are devised that guide development in the "right" direction by, e. g., being resistant to hedging, and that score uncertainty is minimized for a given sample size.

Linkages:

There should be a link with WMO/CBS for the exchange of scores, extending them to surface parameters. The WWRP/WGNE Joint Working group on Forecast Verification Research (JWGFVR) is an important forum for international collaboration on all aspects of verification research.

When going to finer scales, global model developers will also benefit from the experience of the meso-scale community who has been working at the kilometer scale. The THORPEX legacy “High-Impact Weather” (HIWeather) project is addressing issues such as the necessity to create appropriate multi-scale analyses and forecasts for an improved description of high-impact weather. There are other WMO-sponsored programs that address particular forecast challenges which are relevant for global prediction systems. In particular, forecasting problems related to Arctic and Antarctic prediction, and linkages with lower latitudes are addressed by the WMO’s Polar Prediction Project. The Sub-seasonal to Seasonal Prediction Project is particularly relevant to advance research on enhancing the use of extended-range forecasts. Improving predictions and predictive skill of severe weather forecast capacity in developing nations is addressed through the WMO’s Severe Weather Forecast Demonstration Project.

Requirements:

There is a high priority requirement to enhance interactions between model developers, scientists working on advanced diagnostics and verification-type activities in order to accelerate the rate of improvement of forecasting systems by providing proper feedback to model developers.

As systems get more integrated/coupled, there will be a strong requirement for different communities to work together to build and evaluate the models of the future. These communities are related to the atmosphere, ocean, land, cryosphere, atmospheric composition, hydrology and this list might further expand in the coming decades.

In order to deliver a global multi-scale analysis and evaluate forecasts, there will be strong requirements for the global exchange of fine-scale information from ground stations or networks of radars/lidars, well beyond what is currently exchanged.

It is evident that the computing resources needed to run the fine-resolution fully coupled systems of the future, including the relevant description of uncertainty at all scales, and the appropriate re-forecasts for calibration/evaluation will be tremendous. Research is needed on the scalability of our codes, tackling the issue of efficiency of next systems on the future High Performance Computing technologies.

CONCLUSION

It is 110 years since Vilhelm Bjerknes first outlined the paradigm that underpins numerical weather prediction (Bjerknes 1904). Global NWP has already been a hugely

successful scientific and technical enterprise with substantial societal benefits arising from the resultant weather forecasts. The scientific, observational and computational advances that have been needed to realise the Bjerknes paradigm show no sign of letting up so that we can be optimistic that further progress can be expected.

So where will this progress come from? Assuming that the computational capacity and capability is available then there would appear to be every reason to suppose that we are heading for kilometre horizontal mesh sizes in global models. At that resolution we can expect some of the physical parameterisations, such as for deep convection, to be no longer necessary so that those physical processes are described explicitly. This will also greatly reduce truncation errors associated with the way the underlying partial differential equations are solved numerically. We can expect to be huge opportunities for improved weather predictions from increasing the number of processes that are represented in our models within the Earth system, such as the oceans, land surface, sea ice, and composition. This implied increase in complexity will require the science behind coupling of the components, including the data assimilation, to be advanced. To go along with the increasing resolution we can expect a further expansion of the number and type of observations shared internationally to initialise models that will come from novel instruments and platforms.

These scientific opportunities have to be matched by the ability to solve the equations efficiently on supercomputers and this is likely to require new ways to code on massively-parallel machines with potentially millions of cores. The energy consumption (watts per flop) will become a significant challenge. The NWP process has to be viewed end to end in a holistic way with the computer being a critical and fundamental part along with the governing laws. In parallel, the fine-scale information produced in real-time will have to be provided to the users with large increases in data volumes. Data handling and dissemination will be revised to accommodate this new situation. Compression of the information might be needed to communicate only the relevant signal to the users. How will this come about? The key is to think holistically and address the scalability challenge by using novel mathematical solutions and computing techniques. It is no longer good enough to only focus on individual parts of the process.

What can we expect from the forecasts themselves? They will continue to be fundamentally ensemble based with a prediction that defines a most likely state and the confidence one has in that being quantified via a set of scenarios. They are likely to be seamless in the sense that the same model is used in a more-or-less continuous way over a full range of time-scales out to a year or more ahead. And we can expect to be predicting not just the weather but also many other aspects of the atmospheric, oceanic and land-surface environment. Indeed, the use of global models to provide a consistent and complete analysis and prediction of the key attributes of the Earth-system can be used to prevent loss of life, reduce damage and provide economic opportunities. All of this requires an Earth-system approach that in many respects has been pioneered by the weather science and prediction community. Finally the process has to fully engage with the users of these forecasts because society's needs have to be factored into the way the system is developed.

Acknowledgements

The authors would like to thank presenters at the “Environmental Prediction Systems: global and medium-range aspects” session at the WWOSC2014, and David Richardson and Mark Rodwell (ECMWF) for fruitful discussions.

References

- Bjerknes, V., 1904: The Problem of Weather Forecasting from the Viewpoint of Mechanics and Physics”, *Met. Zeit.* pp. 1-7
- Bougeault, P., Z. Toth, many others, T. M. Hamill, and many others, 2009: The THORPEX Interactive Grand Global Ensemble (TIGGE). *Bull Amer. Meteor. Soc.*, **91**, 1059-1072.
- Ferro, Christopher A. T., David B. Stephenson, 2011: Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699–713. doi: <http://dx.doi.org/10.1175/WAF-D-10-05030.1>
- Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and T. N. Palmer, 2012: Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart J. Royal Meteor Soc.*, **138**, 1814-1827.
- Hamill, T. M., 2012: Verification of TIGGE multi-model and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous US. *Mon. Wea. Rev.*, **140**, 2232-2252.
- Hamill, T. M., G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast data set. *Bull Amer. Meteor. Soc.*, **94**, 1553-1565.
- Hirschberg, P.A., E. Abrams. A. Bleistein, W. Bua, L. Delle Monache, T. W. Dulong, J. E. Gaynor, B. Glahn, T. M. Hamill, J. A. Hansen, D. C. Hilderbrand, R. N. Hoffman, B. H. Morrow, B. Philips, J. Sokich, N. Stuart, 2011: A weather and climate enterprise strategic implementation plan for generating and communicating forecast uncertainty information. *Bull. Amer. Meteor. Soc.*, **92**, 1651-1666.
- Holloway, C. E., Woolnough, S. J. and Lister, G. M. S. , 2013: The effects of explicit versus parameterized convection on the MJO in a large-domain high-resolution tropical case study. Part I: Characterization of large-scale organization and propagation. *Journal of the Atmospheric Sciences*, 70 (5). pp. 1342-1369. ISSN 1520-0469 doi: [10.1175/JAS-D-12-0227.1](http://dx.doi.org/10.1175/JAS-D-12-0227.1)
- Hoskins, B. J., 2012: The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Q. J. R. Meteorol. Soc.*, DOI:10.1002/qj.1991
- Lorenz, E. N., 1993: *The Essence of Chaos*. University of Washington Press, 227 pp.
- Magnusson, Linus, Jean-Raymond Bidlot, Simon T. K. Lang, Alan Thorpe, Nils Wedi, Munehiko Yamaguchi, 2014: Evaluation of medium-range forecasts for hurricane

sandy. *Mon. Wea. Rev.*, **142**, 1962–1981. doi:
<http://dx.doi.org/10.1175/MWR-D-13-00228.1>

Rodwell, Mark J, and Coauthors, 2013: Characteristics of Occasional Poor Medium-Range Weather Forecasts for Europe. *Bull. Amer. Meteor. Soc.*, **94**, 1393–1405. doi: <http://dx.doi.org/10.1175/BAMS-D-12-00099.1>

Roff, G., D. W. J. Thompson, and H. Hendon, 2011: Does increasing model stratospheric resolution improve extended-range forecast skill? *Geophys. Res. Lett.*, **38**, L05809.

Shaw, T. A. and T. G. Shepherd, 2008: Atmospheric science: Raising the roof. *Nature Geoscience*, **1**, 12–13.

Smith, G. C., and Coauthors, 2014: Sea ice forecast verification in the Canadian global ice ocean prediction system. Submitted to *Q. J. R. Meteorol. Soc.*

Tripathi, O., and Coauthors, 2014: Review: The Predictability of the Extra-tropical Stratosphere on monthly timescales and its Impact on the Skill of Tropospheric Forecasts. *Q. J. R. Meteorol. Soc.*, DOI: 10.1002/qj.2432.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73-83.

Tables and Figures

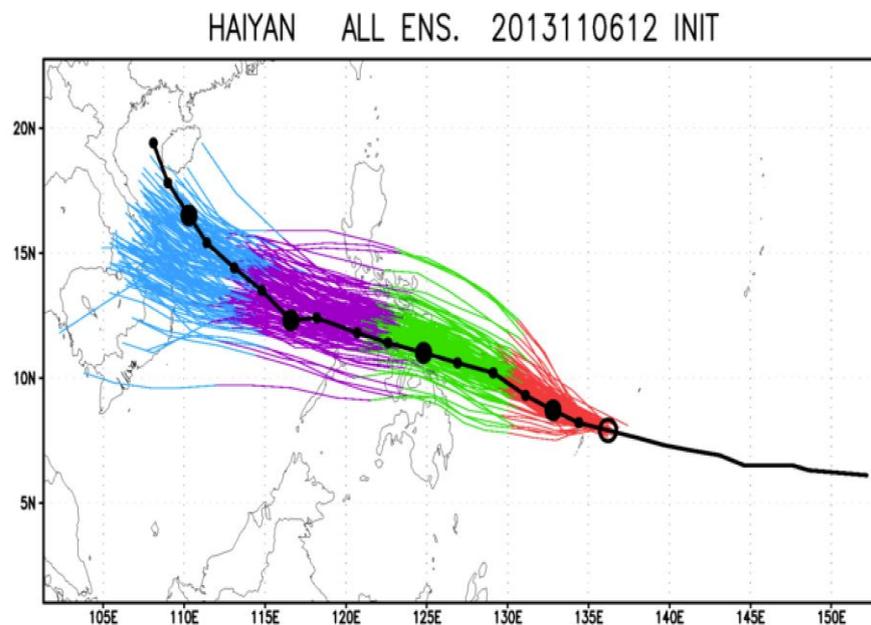


Figure 1: Illustration of multi-model track forecasts using models from the TIGGE data set, here for forecasts of Typhoon Haiyan, initialized at 12 UTC on 06 Nov 2013.

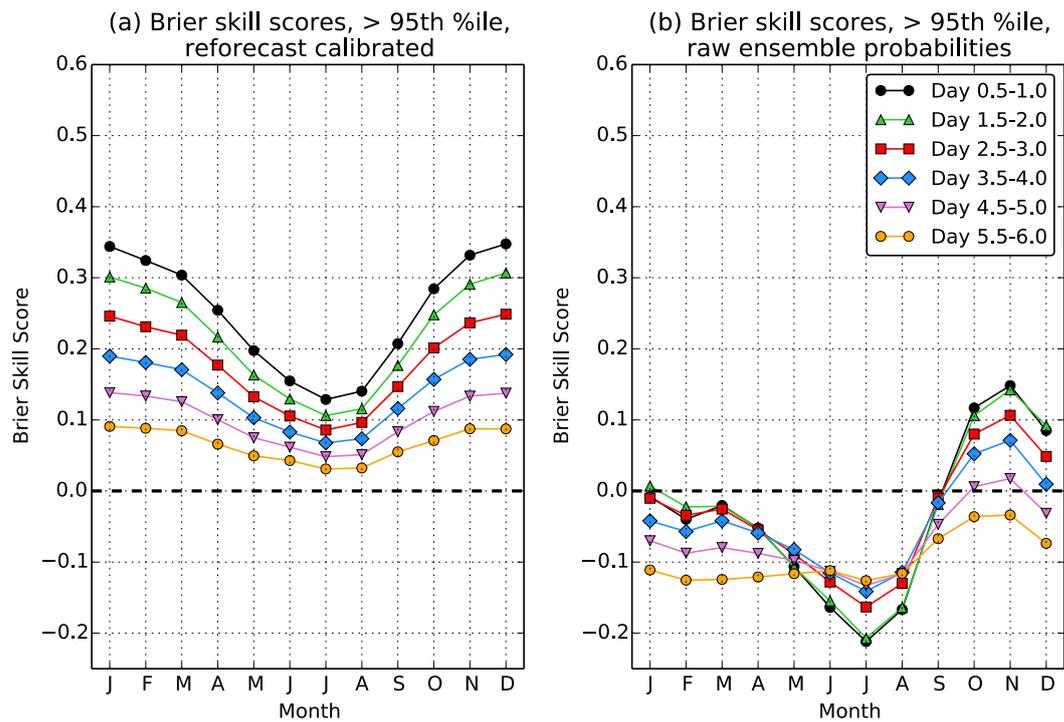


Figure 2. An example of the increased skill provided by post-processing numerical guidance using reforecasts. Brier skill scores for exceeding the 95th percentile of the climatological distribution are shown, calibrated and validated using 2002-2013 1/8-degree precipitation data over the CONUS. (a) reforecast analog-calibrated probabilities using the second-generation global ensemble reforecast data set of Hamill et al. (2013), and (b) raw ensemble probabilities from the 11-member US global ensemble.

N.Hem Extratropics 500 hPa geopotential height

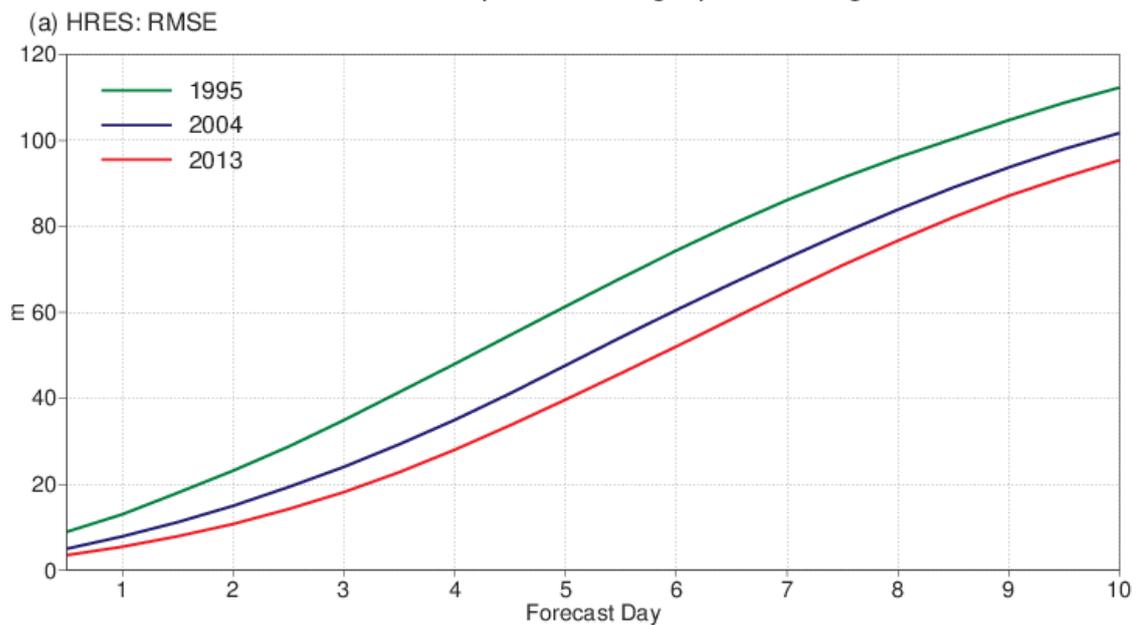


Figure 3: ECMWF's forecast Z500hPa extra-tropical error growth over the last two decades.

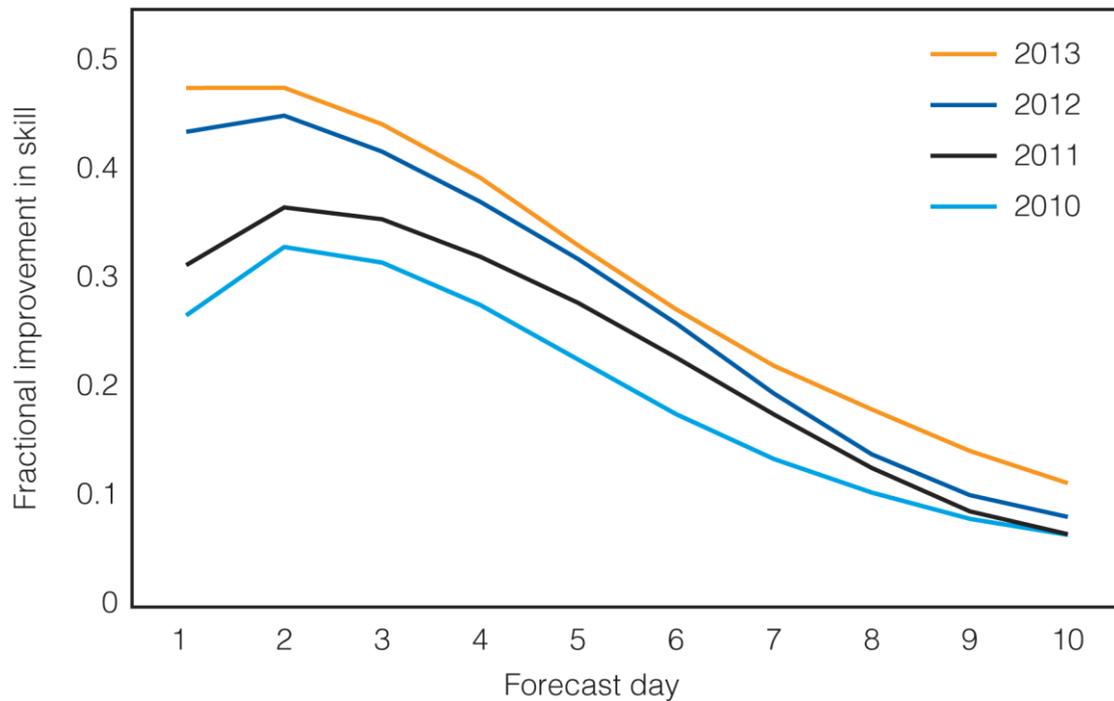


Figure 4: Fractional improvement in the anomaly correlation coefficient at 500 hPa in the extratropical northern hemisphere for the ECMWF high-resolution forecasts compared with those made using the forecasting system of 2006 (ERA-Interim) for calendar years 2010, 2011, 2012 and 2013.

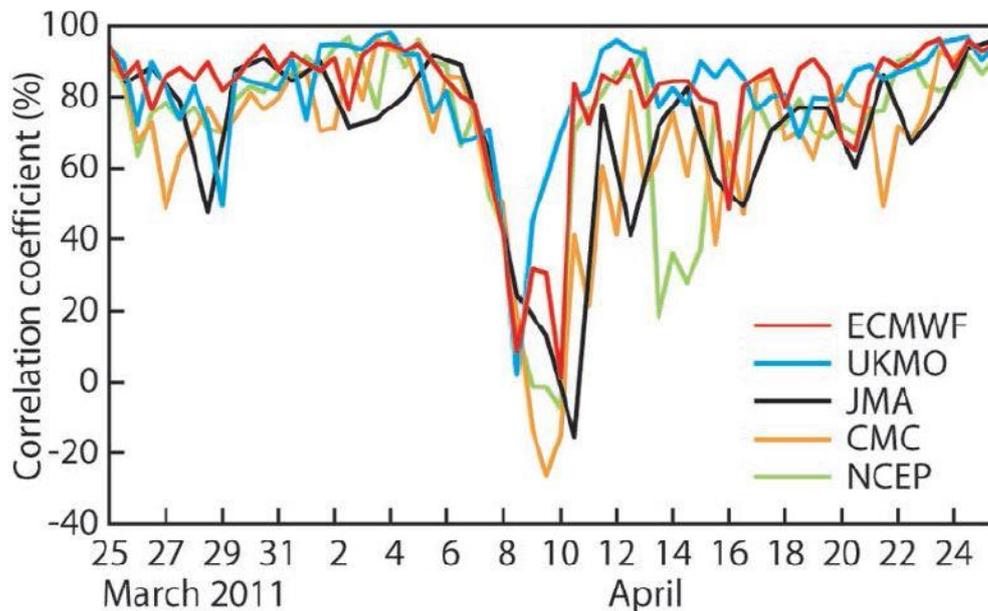


Figure 5: Time series of day 6 forecast skill over Europe from some of the world's NWP centres (within the TIGGE program): Met Office (UKMO), Japan Meteorological Agency (JMA), Canadian Meteorological Centre (CMC), and NCEP. The dates correspond to the start of the forecast. The score shown is the spatial ACC of Z500. Europe is defined in this article as the region 35°–75°N, 12.5°W–42.5°E.