TECHNICAL MEMORANDUM

759

New Developments in the Diagnosis and Verification of High-Impact Weather Forecasts

Mark J. Rodwell[1], Laura Ferranti[1], Thomas Haiden[1], Linus Magnusson[1], Jean Bidlot[1], Niels Bormann[1], Mohamed Dahoui[1], Giovanna De Chiara[1], Sinead Duffy[1], Richard Forbes[1], Elias Hólm[1], Bruce Ingleby[1], Martin Janousek[1], Simon T.K. Lang[1], Kristian Mogensen[1], Fernando Prates[1], Florence Rabier[1], David S. Richardson[1], Ivan Tsonevsky[1], Frederic Vitart[1], and Munehiko Yamaguchi[2]

1) ECMWF, Reading, UK

2) MRI, Tsukuba, Japan

Forecast and Research Department

November 2015

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Abstract

High-impact weather is of particular interest to forecast users, and its prediction with a seamless ensemble analysis/forecast system is at the heart of ECMWF's proposed new strategy. Whilst the impacts of weather are user-specific, from a meteorological viewpoint weather can be high-impact by virtue of its extreme amplitude (e.g. tropical cyclones) or its longevity (e.g. heat-waves). Both aspects present a particular challenge for forecast systems, and demand considerable development of diagnostic and verification capabilities. Here, with reference to numerous high-impact weather types, and particular focus on ensemble aspects, we discuss recent developments in these diagnostic and verification capabilities, and our future plans.

1 Introduction

A high-impact weather is often associated with events that lie in a tail of a climatological distribution for a particular location, and are thus rare and something that the local society is not routinely accustomed to experiencing. Such events could be extreme in amplitude, such as tropical cyclones, intense winds, or heavy convective precipitation. These features present a particular challenge for data assimilation and forecasting. For example, their intensity can mean that the parametrization of physical processes becomes less accurate, particularly in the linearized model. These features are often also associated with small-scale structures that are difficult to observe and resolve, and with large background departures and more observation rejections. Sometimes these extreme features can lead to enhanced forecast errors and uncertainties downstream, often referred-to as forecast 'busts'. ECMWF's proposed new strategy sets as a goal the provision of valuable forecasts of such extreme features well into the second week. Other events may be rare and high-impact by being prolonged 'regimes', such as droughts, heat-waves or cold-spells. These features are a challenge to define, let-alone predict, because they represent the aggregate of interactions between planetary waves (possibly remotely-forced) and smaller spatio-temporal scale variability. ECMWF aims to provide skilful predictions of such regimes up to three weeks ahead. The setting of these goals, and progress towards achieving them, requires considerable upgrades in our diagnostic tools and verification methods, and so a special-topic paper on developments in the diagnosis and verification of high-impact weather forecasts is particularly timely. Furthermore, these and future developments will form an important contribution to the "High Impact Weather" project (Jones et al., 2014) of the WMO World Weather Research Programme. There are others attributes that can make a weather-type 'high-impact' to society, although we do not investigate these here. For example, a weather type can be high-impact, even if it is not very rare, if society is particularly vulnerable to it – such as with the impact of fog on transportation. In theory, a weather-type could be high-impact by virtue of it being inherently less predictable – so society does not have sufficient forewarning to take mitigating action – although many of the relevant weather features will be extreme in amplitude anyway; for example flash floods.

Sensitivity to initial-state and model uncertainty means that weather forecasting is fundamentally a probabilistic task (Sutton 1954; Thompson 1957; Lorenz 1963). ECMWF's proposed new strategy acknowledges this fact by directing research towards a fully integrated ensemble analysis / forecast system, which is as seamless as possible in leadtime and resolution. A guiding principle for developments in diagnostics and verification must be to facilitate the establishment and improvement of such an integrated system. The verification of probabilistic forecasts with 'proper scores', such as the

Continuous Rank Probability Score, will ensure that the forecast system develops in a way that improves both 'Reliability' and 'Sharpness'. Reliability is an overall measure of the agreement between forecast probabilities and outcome frequencies. Such agreement is important to users because if allows them to set the "correct odds" for a given event, and reduces "jumpiness" in event probabilities between consecutive forecasts. Better Reliability can be achieved through modelling improvements: to the forecast model itself (its bias and its representation of the growth of uncertainty), the modelling of observations (observation operators), and in the modelling of observation errors (representation of bias and random errors). Sharpness is an overall measure of how close forecast probabilities are to 0 or 1. Sharper forecasts are desirable for users because there is less uncertainty in the forecast. Better Sharpness can be achieved by reducing the initial-state uncertainty through increasing the information content of the observations: by increasing the number of independent observations assimilated, or by decreasing the magnitude of observation random error. It is clear that both Reliability and Sharpness are critical to the success of the forecast system.

Key diagnostic and verification aspects to consider include, therefore:

- How well high-impact weather is represented in the observations, including biases and random errors, and their sensitivity to observation density.
- How well high-impact weather is represented in our models, including systematic and stochastic errors, and their sensitivity to model resolution.
- The predictability, current predictive skill, and the user's interpretation of forecast value in high-impact weather situations (observed and/or forecast).

Diagnostic tools and verification software are applied to almost all data assimilation and forecast suites produced using the ECMWF Integrated Forecasting System (IFS), and also to some of those available from other institutions in 'The International Grand Global Ensemble' (TIGGE) and the 'Seasonal-to-Sub-seasonal' (S2S) databases hosted at ECMWF. ECMWF data sources include the High-Resolution 4D variational data assimilation (4D-Var; Rabier et al. 2000) and forecast (HRES), the 'ERA-Interim' re-analysis/forecast suite made with a constant IFS cycle, the 25-member Ensemble of Data Assimilations (EDA; Isaksen et al. 2010a,b) and its unperturbed control, the 50-member ensemble forecast (ENS; Palmer et al. 1993; Molteni et al. 1996) and its unperturbed control, and operational re-forecasts (Hagedorn et al. 2012) made for the purposes of quantifying deficiencies in, and calibrating, the operational cycle – particularly at extended ranges – and in the medium-range for the derivation of the Extreme Forecast Index (EFI; Lalaurette 2003).

The topics and results discussed in this paper reflect current research activities in diagnostics and verification at ECMWF. They address some of the most important high-impact weather phenomena, including tropical cyclones, extratropical extreme convection and winds, and European cold-spells. In section 2, diagnostic aspects are discussed. We start with a diagnostic evaluation of our ability to represent tropical cyclones in our models, and the sensitivity to resolution and coupling with the ocean. We also highlight specific issues for data assimilation in tropical cyclone situations. We then discuss, in the context of intense mesoscale convective systems over North America, the diagnosis of random error growth in the forecast model, and whether this is sufficient to maintain ensemble reliability both locally and downstream over the North Atlantic and Europe. Such convective situations can be associated with difficulties to predict a blocking regime over Europe, and we conclude the section with

a discussion of Euro-Atlantic regimes, their transitions and their utility in stratifying predictability. Section 3 focuses on verification. Here we start with an investigation into the use of high-density observation networks as a means of reducing representativity problems, increasing sample sizes and better quantifying forecast skill for extreme precipitation. We then discuss the use of the 'Extreme Forecast Index' as a means of condensing and calibrating ensemble information in the case of severe extratropical convection. Finally the section discusses, from a user-perspective, the potential economic value of forecasts of extratropical windstorms. The paper concludes with a summary of results and a discussion of future plans.

2 Diagnostics of high-impact weather events

In operational forecasting, it is important that diagnostic tools help the Centre improve its forecast performance; by identifying deficiencies in the current forecast system, and possibly suggesting solutions. These tools need to be able to cope with large datasets if they are to identify residual deficiencies in an ever-improving forecast system. To justify the cost of developing these tools, it is often useful to go a step further and make them flexible pieces of code, which can be applied in a variety of situations, to multiple data sources, and by any researcher. Grouped by the main timescale that they are applied too, these tools include

- Short to medium-range: A model-space tool that can derive process-tendency budgets (Klinker and Sardeshmukh 1992; Rodwell and Palmer 2007), as well as analysis increments, forecast errors, and 'Rossby wave source' diagnostics. There is an observation-space tool that can produce 'EDA reliability budgets' (Rodwell et al. 2015; see later) as well as diagnostics of the HRES and ENS. Both these tools are able to produce composites and make comparisons. Statistical significance testing is available throughout. Other tools focus on scale-dependent error and activity, and Potential Vorticity budgets.
- Extended-range: Regime identification and predictive skill tools (Ferranti et al. 2015), that are routinely applied to extended-range operational and re-forecast suites, and to output from other global forecasting centres as part of the Seasonal-to-Subseasonal (S2S) WMO project.
- Model climate: Tools that derive systematic errors, and metrics of variability (blocking, tropical waves etc.), see, e.g., Jung (2005), Magnusson (2015). These are routinely applied to operational and experimental suites, in coupled and uncoupled mode.

Smaller diagnostic modules are also developed. These can be combined flexibly and used to investigate specific issues and case-studies (Magnusson et al. 2014). They are particularly useful in the real-time diagnosis of the operational forecast (part of the in-house 'Daily Report'), and help strengthen the link from case-studies to more systematic and statistically-rigorous diagnostic studies.

There are, of course, many other diagnostic tools throughout the Centre. Major ones include the 'OBSTAT' tool, which is used to characterise observations and their observation operators, including observation biases (Geer et al. 2010) and bias correction (Auligné and McNally 2007), and to provide information on likely errors through, e.g., the application of consistency diagnostics (Desroziers et al. 2005) to the HRES assimilation (Bormann and Bauer 2010). OBSTAT is vital for the monitoring of temporal stability of observations, and for the automatic detection of sudden changes in observation

quality (Dahoui et al. 2014). The adjoint-based diagnostic 'Forecast Sensitivity approach to Observation Impact' (FSOI; see, e.g., Baker and Daley 2000; Langland and Baker 2004; Cardinali and Buizza 2004) is used to estimate the reduction of the forecast error due to the assimilation of the observations. Cardinali (2009) compares this approach with conventional observing system experiments (OSEs). Lupu et al. (2015) show the beneficial impact of applying the Desroziers method to data channels pre-selected by the forecast sensitivity approach. An example of the use of the OBSTAT and FSOI tools is also given in this paper.

Hence diagnostics activity is very much a collaborative exercise at the Centre with, for example, all tools being brought to bear on our forecast systems within the Centre's quarterly joint Forecast Department/Research Department meetings. A cross-sectional meeting is planned for the autumn to further promote communication and collaboration, and to discuss future developments.

The Centre also has diagnostic links with the UK Met Office 'Global Model Evaluation and Diagnostics' section, the UK 'Process Evaluation Group' (PEG) which focuses on blocking and stormtracks, the German 'Pandowae' research group whose focus has been on high-impact weather (and the upcoming follow-on project 'Waves2Weather'), the 'Atmospheric Dynamics' group at ETH Zurich, the University of Ljubljana in terms of tropical wave diagnostics, and Oxford University in their research proposal on 'understanding and representing atmospheric convection across scales'. We are also represented in the WMO working group on 'Predictability, Dynamics and Ensemble Forecasting' (PDEF) and the WCRP Working Group on Seasonal to Interannual Prediction (WGSIP).

As highlighted in the Introduction, the growing importance of ensemble products is reflected in the development of appropriate diagnostics associated with flow-dependent reliability and skill. Here we discuss the two first moments of the forecast associated with reliability in extreme weather situations: bias (which is most directly assessed in terms of systematic errors in single-forecast mode), and variance (particularly associated with the model's representation of the growth-rates of errors and uncertainties). We conclude with a discussion on regimes; their definitions, transitions and predictive skill.

2.1 Diagnostics of intensity errors for tropical cyclones

Tropical cyclones (TCs) are among the deadliest weather phenomena. They develop over subtropical ocean regions; being also known as hurricanes over the Atlantic and typhoons over the Pacific. TCs give rise to a combination of extreme winds, storm surges, high waves and rainfall. When they make landfall, they can have devastating consequences, with TC Haiyan (2013) as a recent example that led to more than 6000 fatalities over the Philippines. Forecasts of TCs several days in advance are sufficiently skilful these days that authorities are able to prepare society for the predicted event, as with TC Phalin (2013) when half a million people were evacuated from coastal regions of Odisha and Andhra Pradesh, India. Further improvements in TC forecasting will help make such actions more precise, and with fewer false-alarms. As discussed above, key attributes of such forecasts are their statistical reliability and their sharpness. Here we focus on one aspect of the reliability component – that of systematic error.

2.1.1 Observations and forecast data

Observations of TC intensity are a critical input for the data assimilation system; for forecast initialisation and forecast evaluation. In-situ observations (surface and radiosonde measurements from

land stations, ships, buoys) are fairly scarce except at the time of landfall. In the oceans around the U.S, reconnaissance flights provide additional dropsonde measurements. Due to the general lack of in-situ observations, satellite observations are of high importance, such as sea surface winds from scatterometer instruments. Observations available for verification are described in "Verification Methods for Tropical Cyclone Forecasts" (WMO 2013). The position and intensity of tropical cyclones are subjectively assessed by meteorologists at tropical cyclone warning centres using all available observations, and the estimate is put into the "Best Track" database (Knapp et al. 2010; Levinson et al. 2010). A main source of information here are visible and infrared satellite imagery that are used to estimate the intensity of cyclones by the Dvorak technique (Dvorak 1975; Velden et al. 2006). The errors in the intensity estimation using the Dvorak technique in comparison to aircraft measurements was investigated in Martin and Gray (1993) and more recently the uncertainty in the Best Track database was investigated by Landsea and Franklin (2013).

Here we will evaluate operational forecasts from 2013-2014, together with experiments that estimate sensitivity to model resolution and coupling with the ocean. We will also show examples of diagnostic tools applied to the data assimilation. The operational HRES is run without an interactive ocean, at resolution T_L1279 (16 km) with 137 model levels. The horizontal resolution for the ENS is TL639 (32 km). In November 2013 the number of model levels for the ENS increased from 62 to 91 and, at the same time, coupling to the ocean from the start of the forecast was introduced (Janssen et al. 2013).

2.1.2 Intensity errors during 2013-2014

The performance of tropical cyclone forecasts is continuously monitored in our standard evaluation (Haiden et al. 2014). Focusing on the 2013-2014 period, Figure 1a shows the mean intensity error as a function of lead time for HRES (red) and the ENS Control (black), evaluated for all tropical cyclone basins. The tropical cyclone tracks are determined by the tracker described in Vitart et al. (2012). To avoid the introduction of artificial biases in the results, a cyclone is only included in the verification if it exists, at the time of forecast initialisation, both in the observational Best track database and in the feature-tracking algorithm applied to the forecast. As a consequence, there is a reduced sample size for longer lead times. For example, the sample size at day 2 is 462 cyclones for HRES while at day 6 the sample size is reduced to 54 cases. The mean intensity error for short forecasts is positive on average (too weak cyclones). The error is present already at initial time, indicating that the data assimilation does not reproduce the minimum pressure in Best Track. However, for longer lead times there is an indication that a negative bias is developing (too strong cyclones) in the HRES forecast. Although the sample is relatively small, the negative bias is statistically significant (at the 5% level) for day 7 and 8 (168h and 192h).

In order to investigate the intensity bias further, Figure 1b shows the error in minimum pressure for day 2 HRES forecasts in tropical cyclones compared to the Best Track analysis for the north-western Pacific. Averaged over all these TCs, the minimum pressure is too high (cyclones too weak), as expected from the results above (Figure 1a). However, there appears to be a spatial pattern to the errors, with cyclones in the southern part of the domain having a positive minimum pressure error, while in the north-western part of the basin (south of Japan) a majority of the cyclones have a negative error. This pattern agrees with the bias in the JMA global model.

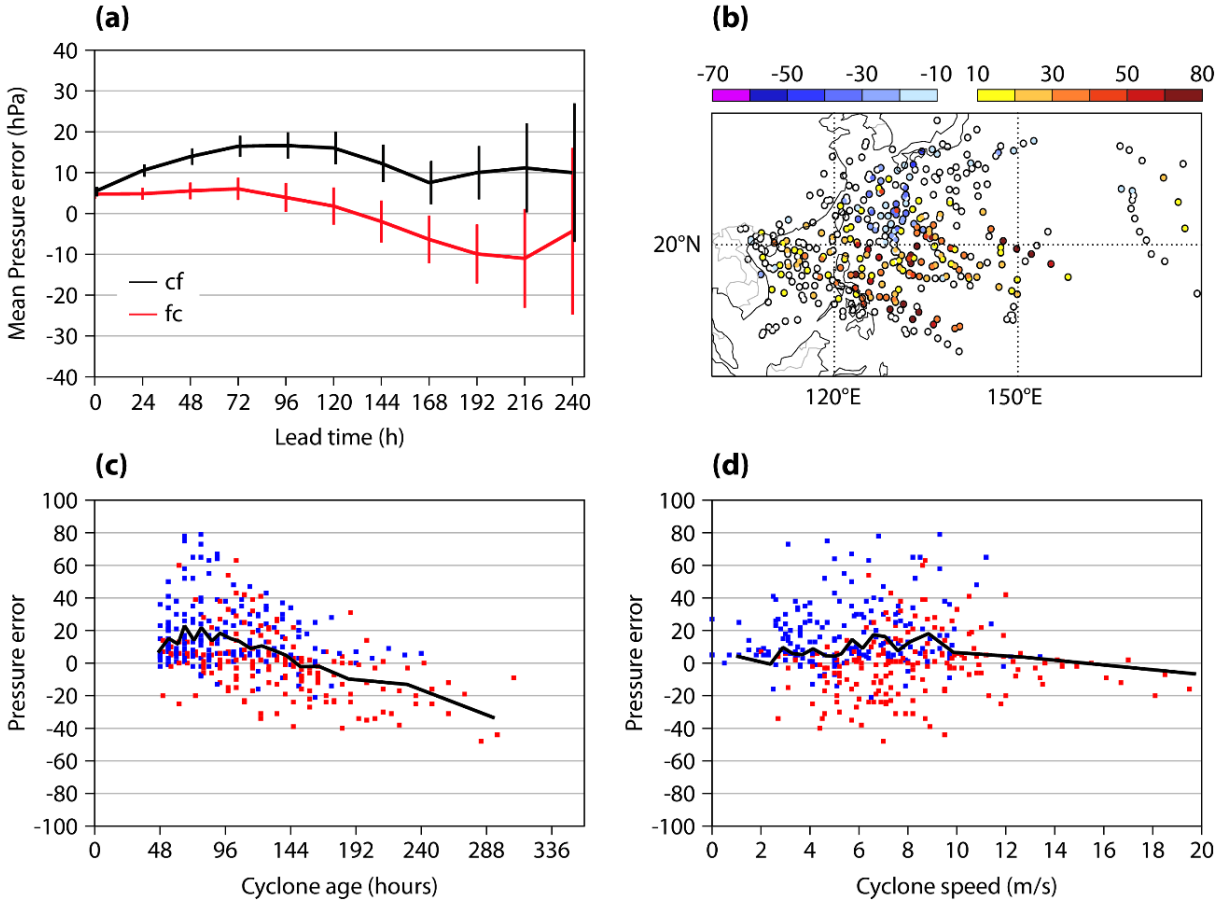

Figure 1: (a) Mean intensity error (in terms of minimum pressure) as a function of leadtime for 2013-2014 from HRES (red) and ENS Control (black) for all tropical cyclone basins. (b) Day 2 error in minimum pressure for all verified tropical cyclones 2013-2014 over the NW Pacific. (c) Error in minimum pressure as function of cyclone age for day 2 forecasts over the NW Pacific. (d) Error in minimum pressure as function of cyclone propagation speed for day 2 forecasts over the NW Pacific. Cyclones south of 20°N (blue) and north of 20°N (red).

Figure 1c shows the error of the minimum pressure in day 2 forecasts as a function of cyclone age, measured in hours from the identification as a tropical storm in Best Track. The blue symbols are for cyclones south of 20°N and the red north of 20°N. The black line represents the average error as a function of cyclone age (the averaging involves binning on cyclone age, with each bin containing 20 cyclones). As we restrict the verification to cases where the cyclone was classified as at least a tropical depression at initial forecast time, the minimum age for day 2 forecasts is 2 days. The error is in general positive for younger cyclones that are located in the southern part of the basin. According to the bin average the sign of the error changes for cyclones older than 7 days. This suggests that the intensity bias change at day 7, seen in Figure 1a, is partly associated with the age of the tropical cyclone and not simply the forecast leadtime. By inspecting individual forecasts the error for some of them partly indicate phase errors in the intensity with a too slow intensification and a delayed weakening. In many cases the error appears already in the initial conditions, as the analysis is to a large extent a product of the background forecast in sparsely observed regions.

Figure 1d shows the intensity error as a function of the cyclone propagation speed. The group of very fast cyclones (>14 ms^{-1}) are mainly cases undergoing extra-tropical transition. The results indicate that there is no strong relation between the intensity error and the propagation speed. This could be explained by the fact that there are only few cyclones with very low (<2 ms^{-1}) propagation speed. In the idealised study by Halliwell et al. (2015) the large sensitivity to propagation speed was found for slow-moving cyclones over relatively shallow warm-layer ocean. For cyclones moving faster than 4 ms^{-1} the sensitivity was small. Errors might be more associated with the TC size, and the depth of the ocean mixed layer.

To summarize the results for HRES intensity errors during 2013-2014, we find a positive (too weak) bias for short (1-5 day) forecasts, while longer forecasts indicated a tendency for too strong cyclones. Too strong cyclones are mainly located in the north-western part of the basin. By plotting the intensity error as a function of cyclone age for the north-west Pacific basin we found that also short-range forecasts (2-day) are too intense for old cyclones (~6 days after cyclogenesis). However, is it so far not clear which physical processes (air-sea interaction, dynamics, convection, microphysics, boundary layer parameterisations etc.) cause the cyclones to deepen too strongly. For the control forecast, the mean intensity error is positive for all lead times. This could be related to the resolution and/or the coupling to the ocean (see section 3.1.4).

2.1.3 Data assimilation aspects

Assimilation of tropical cyclones is challenging for several reasons. For example, conventional observations are sparse in tropical cyclone regions, and most satellite radiances are rejected due to cloud contamination. In the vicinity of the cyclone centre, gradients are sharp and a small error in the cyclone's position in the background will give rise to large departures from the observations. One approach to improve the understanding of the data assimilation method is to investigate case studies with diagnostic tools that can target and follow a tropical cyclone. In this section we will give examples of such diagnostics used to understand the assimilation of TC Gonzalo (2014). Gonzalo formed east of the Antigua and struck close to Puerto Rico before turning northward and later hitting Bermuda. Observations near the tropical cyclone were available intermittently (mainly close to islands), with a gap of observations near the tropical cyclone on 16 and 17 October when the cyclone was between Puerto Rico and Bermuda. During the second half of 15 October the observed pressure was much lower than the first guess and analysis. This was then interpreted as a surface pressure observation bias and later observations were wrongly bias corrected - this problem is now under further investigation.

Figure 2 shows observation statistics for the meridional wind component from dropsondes in the lower part of the troposphere (600 hPa to surface), and the surface pressure field, for the data assimilation cycle 14 October 06UTC. Figure 2a shows observation-minus-background (coloured circles) and the background surface pressure field, Figure 2b shows observation-minus-analysis and the analysis surface pressure field, and Figure 2c shows the Forecast Sensitivity Observation Impact (FSOI, Cardinali 2009) and the same analysis surface pressure field. The black triangle represents the cyclone position reported in the Best Track data set. In the background the cyclone was located north-west of the reported position (Figure 2a) and consequently there were large differences with observations. The pressure field in the analysis hardly resembles a tropical cyclone, although the observations fit the analysis better (-12.6 versus -0.8 ms^{-1} in mean). The FSOI values are positive for all but one dropsonde meaning that the

observations contributed to an *increased* 1-day error. This issue was probably not a result of erroneous observations, but rather a combined result of the large background position error, low inner-loop resolutions in the data assimilation, and the concentration of dropsondes on one side of the tropical cyclone. The results agree with Harnisch and Weissmann (2010) who found little improvement from dropsondes if they were only deployed close to the cyclone centre. Cases where dropsondes degraded the analysis of tropical cyclones were highlighted in Aberson (2008).

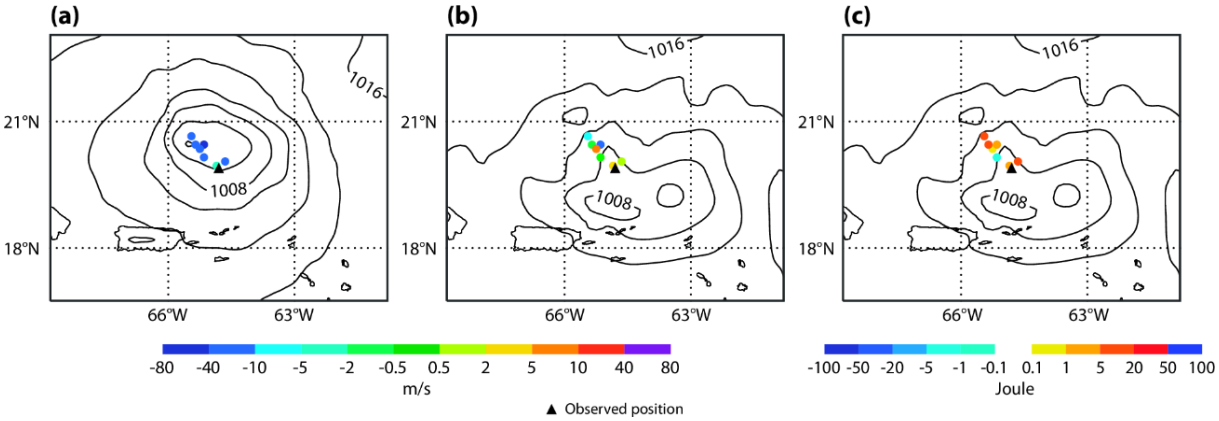

Figure 2: (a) Background departures, (b) Analysis departures and (c) Forecast Sensitivity Observation Impact (FSOI) values, for dropsonde wind observations in the lower troposphere. Surface pressure fields are included in the panels from (a) the background and (b and c) the analysis.

One of the main sources of information about tropical cyclones available to the data assimilation is scatterometer data, measuring surface wind speed and direction. However, due to the measurement method, the direction has a 180° ambiguity which is resolved by comparison with the background (and subsequent outer-loop) wind field. A wrong background wind field can lead to the wrong selection of wind direction. Another issue with the way ECMWF uses the scatterometer observations is that the number of wind vectors used is much less than the number available. Before the start of the assimilation, wind vectors are thinned in order to avoid too large weights being assigned to the observations, and to take into account spatial correlations in observations error. Scatterometer winds are not available above 35 ms^{-1}, due to backscatter saturation effects. In addition to this, some of the observations close to the centre of the TC are rejected by the quality control due to large misfits (of the position of the storm). Ongoing research is addressing these difficulties.

2.1.4 Sensitivity to horizontal resolution and ocean coupling

High model resolution is required to resolve the processes that are active within a tropical cyclone. At present (2015), limited area models for tropical cyclones typically use a resolution of a few kilometres (Davis et al. 2008; Gopalakrishnan et al. 2011; Gall et al. 2013). In a global model with resolution of O(10km), only a few grid points cover the core of the cyclone. In this section we test four different resolutions: TL639 (32 km), TL1279 (16 km), TCo639 (17 km) and TCo1279 (9 km). The “L” in “TL” relates to the “linear” grid. “Co” in “TCo” relates to the soon-to-be implemented cubic octahedral grid – which has about twice the grid resolution for the same spectral resolution (Wedi et al. 2015), and this allows the model to have almost no explicit spectral diffusion in the troposphere (personal

communication Sylvie Malardel, 2015). All simulations in this section use 137 vertical levels and are initialised from the same analysis (based on TL1279).

Figure 3 shows the evolution of the minimum pressure for 2 initial dates for (a) TC Haiyan from 2013 (left) and (b) TC Neoguri from 2014 for the four resolutions (TL639 - red, TCo639 – light-blue, TL1279 - blue, TCo1279 - green). The Best Track data is presented in black. Different line styles represent the different initial dates. For Haiyan, the intensity in general is too weak in the forecasts, and the increased resolution clearly improves the intensity, especially between using TL1279 and TCo1279. For Neoguri, the forecast is too weak in the first phase of the cyclone. In the late stage, when the TL1279 forecasts are too intense, the error is increased further with higher resolution (TCo1279).

Figure 3 (c) shows the difference in minimum pressure averaged over 21 forecasts for 8 different tropical cyclones in the north-western Pacific during 2013 and 2014, relative to the TL1279 resolution simulations. The statistics are limited to the first 96 hours to maximise the sample size. This set of cases confirms the results from the individual cyclones in the top panels (which are included in the sample). The TL639 minimum pressure is on average 10 to 15 hPa higher in the 48 to 96 hour forecast range compared to TL1279 (red), while the TCo1279 simulations produce about 10 hPa deeper cyclones than TL1279 (green). However, the difference between TCo639 and TL1279 is very small (blue), indicating that the cubic octahedral grid is more computationally efficient in the modelling of tropical cyclones.

A key energy source for tropical cyclones is the heat flux from the ocean. By not coupling the atmosphere and the ocean, there is no feedback from the heat exchange at the surface and the ocean acts as an infinite source of energy for the atmosphere during the forecast. Using a coupled model introduces a negative feedback between the tropical cyclone and the sea-surface temperature (SST), especially for slow-moving cyclones or cyclones over a shallow warm ocean layer. The winds from the cyclone increase the heat flux from the ocean that cool the SST, thus reducing the heat flux in the next model step and the energy supply to the cyclone. Tropical cyclones interact with the SST in three ways: (1) the heat flux to the atmosphere (enhanced by high wind speeds) which cools the surface, (2) the wind speed increases the vertical mixing and (3) upwelling by Ekman pumping. The first process could be simulated by a slab ocean model, the second requires a mixed-layer model and to simulate all three processes a 3-dimensional model is required, which is important especially for slow moving cyclones (Yablonsky and Ginis 2009). The effect of a mixed-layer model in ECMWF low-resolution forecasts (TL159) for tropical cyclones was investigated in Takaya et al. (2010), however the low model resolution led to a negative intensity bias and all cyclones were relatively weak.

We now show results from preliminary experimentation with a high-resolution atmosphere (TCo1279) coupled to the Nemo ocean model with resolution 0.25°x0.25°. Currently the high-resolution (0.25°x0.25°) ocean reanalysis (Zuo et al. 2015) is available only for a limited number of years and every 5th day, which restricts the set of initial dates for this investigation. In the uncoupled simulations with persisted SST anomalies, the forecast is initialised with SST from the ocean reanalysis (instead of OSTIA SST as operational uncoupled forecasts), in order to have the same initial SST in both coupled and uncoupled mode.

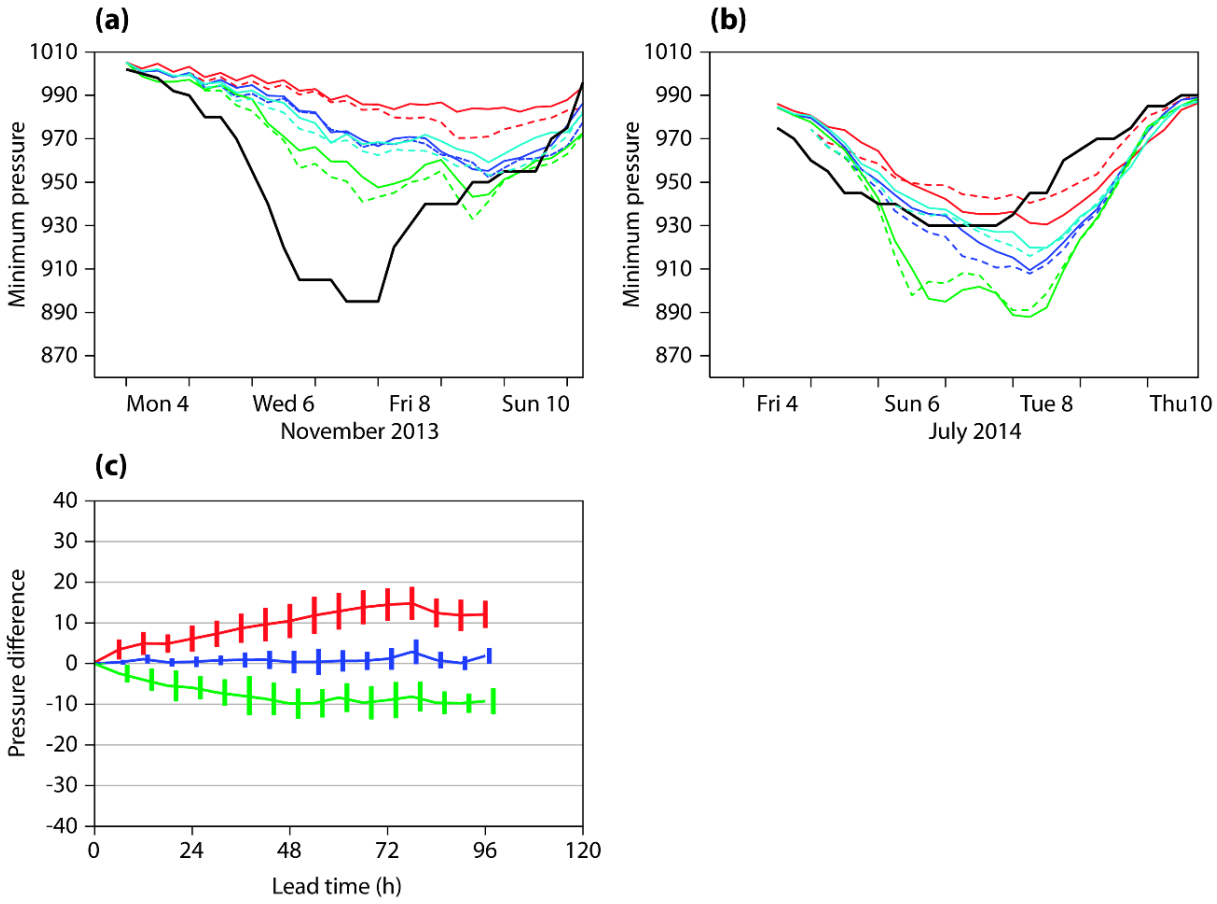

Figure 3: (a) Minimum pressure forecasts for TC Haiyan, initialised at 0UTC on 4 November (solid) and 5 November (dashed) 2013. (b) Minimum pressure forecasts for TC Neoguri initialised at 0UTC on 4 July (solid) and 5 July (dashed) 2014. In (a) and (b), the different resolutions are represented by (red) TL639, (light blue) TCo639, (dark blue) TL1279 and (green) TCo1279. The Best Track pressures are given in black. (c) The difference in minimum pressure forecasts, relative to TL1279, based on 21 forecasts of 8 tropical cyclones during 2013 and 2014: (red) TL639-TL1279, (blue) TCo639-TL1279 and (green) TCo1279-TL1279.

Figure 4 (a and b) show a comparison between uncoupled (blue) and coupled (red) forecasts of the evolution of the minimum pressure for (a) Haiyan and (b) Neoguri. As seen above, Haiyan represents a tropical cyclone where the intensity was under-estimated and Neoguri represents a case where the intensity was over-estimated in its latter stage. While the effect of the coupling is very small for Haiyan (Figure 4a), the intensity is much reduced in the Neoguri case (Figure 4b). The small effect of the coupling for Haiyan could be connected to the deep and well developed ocean mixed-layer for this case (Lin et al. 2014).

Focusing on the effect of the coupling for Neoguri, Figure 4 (c) shows the SST for 6 July +96h in a run with coupled model, together with the observations of SST from 10 July shown with diamonds. Comparing SSTs within the wake of the TC with SSTs outside the wake, the effect of the coupling reaches 5°C in the wake of the cyclone, which is in line with studies for other cyclones in the literature (e.g. Halliwell et al. 2015). In the central part of the wake we have two observations that are in close

agreement with the SST in the coupled run. Figure 4(d) shows the time series of the southernmost of these two observations and forecast values (both coupled and uncoupled) for the observation location. The evolution of the SST in the coupled forecast agrees well with the observation, which also shows a cooling of 5°C over 24 hours.

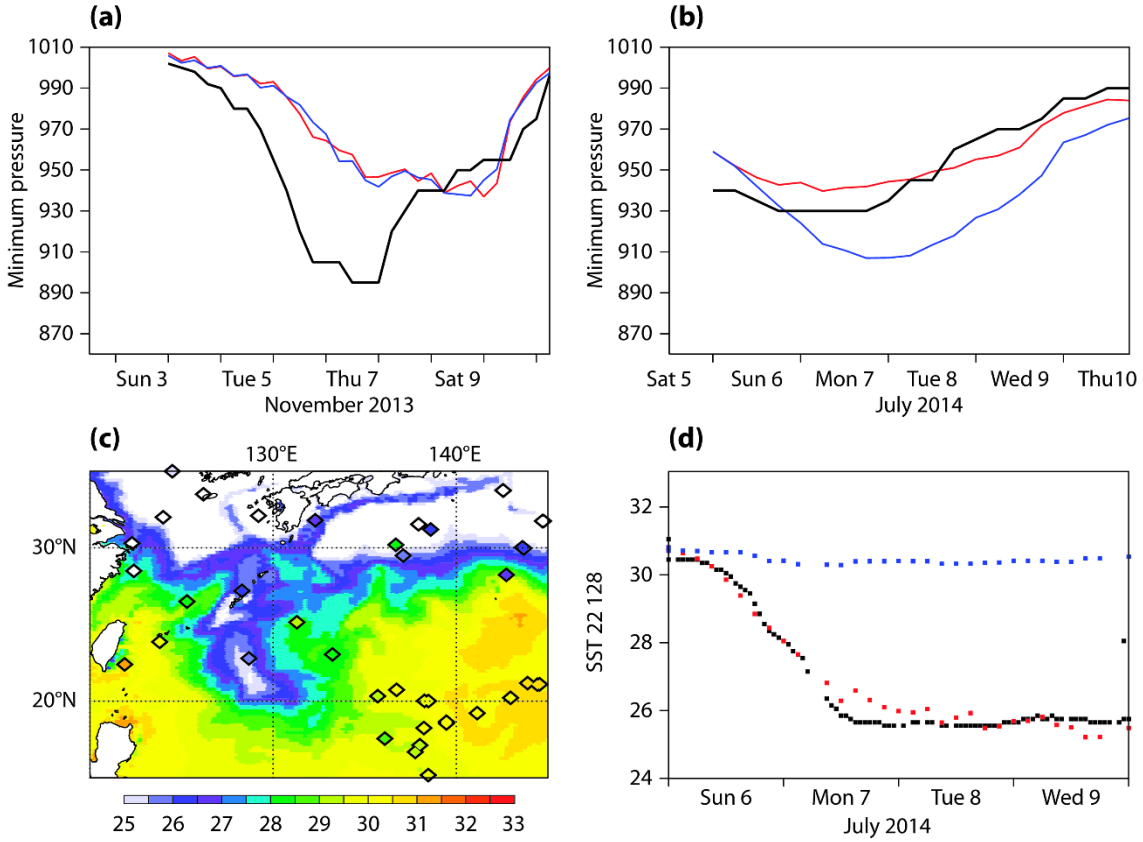

Figure 4: Central pressure forecasts for (a) TC Haiyan and (b) TC Neoguri, with TCo1279 resolution in the atmosphere: (blue) uncoupled model and (red) coupled model. (c) Shading shows D+4 SSTs from the coupled forecast initialised at 0UTC on 6 July 2014. Diamonds show the contemporaneous observations. (d) SST time-series from the (black) observations, (blue) uncoupled forecast and (red) coupled forecast at the location of a buoy with approximate position 22°N, 128°E.

Our results qualitatively agree with the findings in Ito et al. (2014) where forecasts for 34 tropical cyclones south of Japan were compared from the operational JMA model with 20 km resolution and a non-hydrostatic limited-area model with resolution of 5 km. The intensity error increased (too intense) by using the 5 km model compared to 20 km but was improved by using the coupled high-resolution model. Hence both higher resolution and coupling are needed.

2.1.5 Summary

Forecasting the intensity of tropical cyclones is still a challenge in numerical weather prediction, both from modelling and data assimilation aspects. Although the dominant intensity error is still too weak cyclones, especially in the analysis and in the early forecast ranges, several cyclones in recent years have been developing too deeply in the HRES forecasts, most frequently when approaching Japan. Intensity errors originate both from the initial conditions and the model. In this section we have shown

examples of tailored diagnostics for tropical cyclones in the context of observation usage in the data assimilation, sensitivity to model resolution and air-sea coupling in the forecasts.

The availability of observations around tropical cyclones is a problem both for data assimilation and verification. For verification we generally use the Best Track estimate that, for many cases, is dominated by the use of Dvorak technique were the intensity is determined from pattern in satellite images. In data assimilation, one way to increase the information going into the data assimilation is to also assimilate the Best Track position and intensity, as currently done at the Met Office (Julian Hemming, personal communication). A similar approach is used at JMA, while NCEP uses a vortex relocation scheme (Hsiao et al. 2010). However, the intensity estimate is not perfect and a good estimate of its errors is considered necessary before it can be assimilated at the Centre. The performance also needs to be carefully evaluated away from the tropical cyclone. For data assimilation with 4DVAR, the resolution of the increments (inner-loop resolution) and the linear assumption in the scheme are also key aspects, however these have not been discussed in this report and require further diagnosis.

Studying the leadtime dependence of intensity bias, we found in general too weak cyclones for short lead times and for HRES a drift towards too intense cyclones for long lead times. However, the bias for long lead times is connected to the issue that the verification here only samples mature cyclones and this could influence the results. Preliminary results from coupling TCo1279 to the 0.25° Nemo model suggests that the over-deepening is at least partly due to lack of coupling to the ocean. For young cyclones, which are usually smaller in horizontal scale, the model resolution is still a limiting factor, and with increased model resolution the intensity of these cyclones will improve.

There could be other factors that limit the ability to model rapid intensification and weakening of tropical cyclones that need to be explored. We are using model tendencies to diagnose the physical processes within tropical cyclones, although this has not been illustrated in this report. We have focused here on the minimum pressure as a measure of intensity. Further diagnostics are required to assess the relationship between pressure gradients and maximum wind speeds. In addition, the forecasts of the TC tracks and interactions with the environmental flow require more diagnostic work. It is known that the extratropical transition of TCs can lead to the development of large amplitudes in the extratropical planetary waves, baroclinic development and further high-impact weather (Jones et el. 2003). Such extratropical transitions are also beginning to receive more diagnostic attention.

The investigation here has focused on how we might reduce systematic errors for tropical cyclones. For a reliable ensemble system, it is also important to represent the random aspect of forecast error (*i.e.* forecast uncertainty). It is becoming recognised that the extratropical transition of tropical cyclones can lead to increased forecast uncertainty (including over Europe) but, owing to the variable location of these transitions, diagnosis of random error growth is not straight-forward. In the next section, we show developments in the diagnosis of forecast uncertainty but, instead of tropical cyclones, we focus on mesoscale convection over North America which is known to increase forecast uncertainty and, owing to its more fixed location, is more amenable to diagnosis.

2.2 Diagnostics of ensemble reliability following mesoscale convection

Mesoscale convective systems (MCSs) are regions (~500km in scale) with embedded intense convective cells (~40km in scale) which can lead to flooding, and embedded tornados which can lead to wind

damage. These scales of convection and interaction with the large-scale make MCSs hard to represent well in current global models, especially ensemble simulations (the EDA outer-loop resolution is ~50km). It is possible that global forecast models will continue to have difficulties representing MCSs until they are able to adequately resolve the embedded convective cells. Over North America, MCS events can be associated with a trough anomaly over the Rockies and, ahead of the trough, northerly advection of warm, moist air that leads to high values of Convective Available Potential Energy (CAPE). Following case-study work by Grazzini and Isaksen (2002), Rodwell et al. (2013) highlighted a statistically-significant link between such 'trough/CAPE' situations in the initial conditions and HRES forecast 'busts' at a leadtime of 6 days over Europe (and associated with uncertainties in the prediction of the onset of blocking; Mauritsen and Källén 2004). Errors in the strong interaction between the jet-stream and the MCS out-flow (at ~200hPa) were seen as the crucial link. The ENS at D+6 also highlighted increased forecast uncertainty (*i.e.* larger ensemble variance or "spread"), associated with such initial conditions (*c.f.* 7a and b of Rodwell et al. 2013). However, there was an indication (not statistically significant) that the ENS did not display a sufficient increase in spread to reflect the errors that occurred (based on a composite of 84 trough/CAPE events within the period 20101110 - 20120320). Such a deficiency, if real, would imply a lack of ensemble reliability for these flow situations. Hence there are several reasons to re-visit MCSs over North America, in order to better determine their role in error-growth rates and ENS reliability.

Before presenting results, it is worth first considering a little further what is meant by ensemble reliability. Figure 5 (re-drawn from Rodwell et al. 2015) shows a perspective diagram of a reliable ensemble forecast system. The light blue curves depict the trajectories of individual ensemble members in (2D; "x,y") state-space as a function of leadtime ("t", coming out of the plane). At any given leadtime, the ensemble members represent independent samplings of an underlying distribution (depicted by the light-blue shaded ellipses; and can be thought-of as the distribution one would get from an infinite ensemble). For a reliable forecast system, the truth (black curve) also represents a sampling of the same distributions (Hamill 2001; Saetra et al. 2004). For medium-range forecasts ($t \geq 2$ days), the analysis (pink dot) is often an adequate approximation for the truth (black dot). The right side of the green triangle in Figure 5 represents the error of the ensemble-mean (dark-blue dot) relative to this analysis. The error can also be written in terms of the other two sides of the triangle, which represent the differences of the analysis and ensemble-mean from the distribution-mean (light-grey dot). Using the ensemble and analysis values from a sufficient number (n) of forecasts, the mean-squared lengths of the sides of the triangle can be estimated, and this leads directly to the often-used "spread-error" equality for reliable forecast systems (see, *e.g.*, Leutbecher and Palmer 2008). More generally, this relationship can be written as

$$\text{Error}^2 = \text{EnsVar} + \text{Residual} \quad , \qquad (1)$$

where Error^2 is defined as the mean-squared error of the ensemble-mean, EnsVar is the mean sample variance of the ensemble (scaled by $\frac{m+1}{m-1}$ to take account of the finite size m of the ensemble), and Residual represents any imbalance due to sampling uncertainties or deficiencies in reliability (its expected value is zero for a reliable system). It is important to note that the n forecasts considered could

represent a sub-set of forecasts associated with a particular initial state, and here we will base the analysis on a composite whose initial states project strongly onto the trough/CAPE patterns.

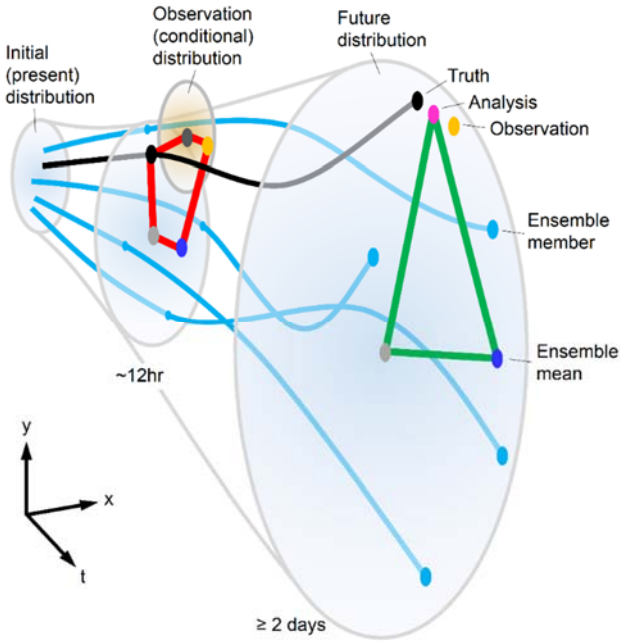

Figure 5: Perspective diagram (time coming out of the plan) highlighting distributions of initial (*i.e.* present) and future uncertainty (light-blue shaded ellipses) and how these are sampled by independent ensemble forecasts (blue curves). For a 'reliable' forecast system, the truth (black curve) should also represent a sampling from the same underlying distributions. The ensemble-mean (dark-blue dots) will not necessarily lie at the centres of these distributions - particularly for small ensemble sizes. In the medium-range ($t \geq 2$ days), the analysis (pink dot) is often taken to be the truth, and the green triangle highlights the basis of the standard 'Spread-Error' relationship associated with reliability. However, non-linear interactions (represented schematically by the crossing of blue curves) tend to prevent the identification of localised deficiencies in error growth-rates. At shorter leadtimes ($t \sim 12$hr) this identification is potentially more straight-forward, although uncertainties in our knowledge of the truth cannot be ignored. This leads, in 'observation space' (the orange dot is an observation), to the inclusion of extra terms associated with bias and random observation error as highlighted by the red pentagon; which forms the basis of the 'EDA reliability budget' discussed here. The additional uncertainty associated with observation error (on top of that associated with the truth) is represented by the conditional distribution, as depicted with the orange shaded ellipse. All distributions depict two state-dimensions, "x" and "y" (or two variables in observation-space) and, for clarity alone, are drawn uni-modal; the central grey dots representing the distribution-means.

Focusing here on the period 20131119 – 20150512, for which IFS cycle 40R1 was operational, 54 trough/CAPE cases in the initial conditions for the EDA control were identified (using exactly the same method as in Rodwell et al. 2013). The number of EDA cycles for which the initial conditions did not fulfil the trough/CAPE criteria was 1019. Figure 6a shows mean 850 hPa specific humidity and horizontal winds in the EDA control analysis for the trough/CAPE composite, and Figure 6b shows the corresponding fields for the non-trough/CAPE composite. This figure (and a similar one for temperature,

not shown) emphasise the anomalous strength of the northward humidity (and temperature) advection within the trough/CAPE composite.

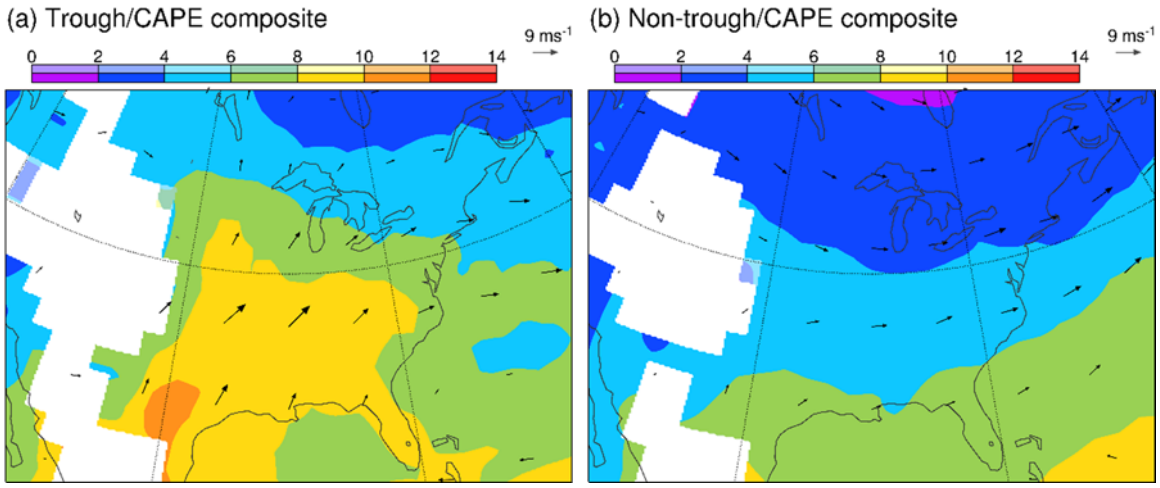

Figure 6: Mean specific humidity (gkg^{-1}) and horizontal winds at 850 hPa for (a) the trough/CAPE composite and (b) the non-trough/CAPE composite.

To examine the precipitation that follows trough/CAPE and non-trough/CAPE events, we use radar-derived precipitation data. These data (for the North American region only) are assimilated at ECMWF as the transformed quantity $\ln(RR6 + 1)$, where RR6 is the "rain-rate" in mm/6hr (Lopez 2014a,b). Figure 7 shows, the composite-means of $\ln(RR6 + 1)$ within the subsequent data-assimilation window (between leadtimes 3 and 15hr). Although the transform will down-weight extreme precipitation values, it is nevertheless clear that the trough/CAPE composite (Figure 7a) does group high precipitation events relative to the non-trough/CAPE composite (Figure 7b). In units of rain-rate, the main precipitation region in Figure 7a displays mean precipitation values at least 40% greater than in the non-trough/CAPE composite.

Figure 8 shows the terms of (1) for 200 hPa geopotential at leadtimes 1, 3, and 5 days for the trough/CAPE composite. Large mean-squared errors ($Error^2$) at D+1 (relative to the ENS control analysis) can be seen over the Great Lakes region of North America (Figure 8a), along the northern edge of the MCS convective region (Figure 7a). These errors are reasonably well predicted, on average, by the ensemble variance (EnsVar, Figure 8b). As the leadtime increases, this signal is seen to propagate east across the North Atlantic (D+3 in Figure 8d,e and D+5 in Figure 8g,h). Notice in particular the large $Error^2$ at D+5 close to western Europe (Figure 8g). This error is more than 15% greater than for the no-trough/CAPE composite (not shown), but the variance is actually slightly smaller than for the no-trough/CAPE composite (not shown). The right panels in Figure 8 show the Residual ($=Error^2 - EnsVar$). The positive residual near western Europe at D+5 (Figure 8i), while not significant at the 5% level (saturated colours indicate statisical significance), is consistent with the ensemble variance being too small (and consistent with the result obtained in Rodwell et al. 2013, from the earlier data period).

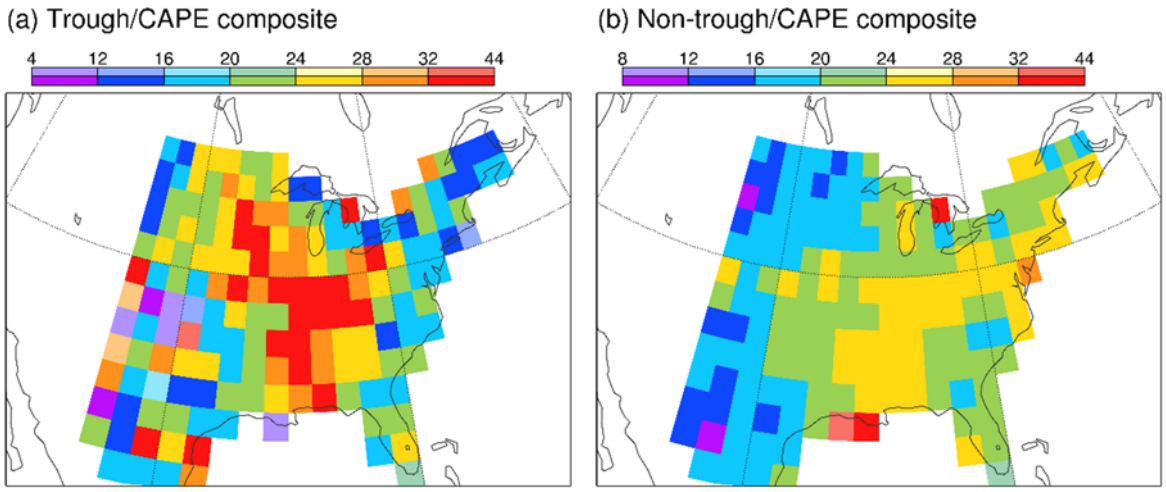

Figure 7: Mean (transformed) precipitation fields for (a) the trough/CAPE composite and (b) the composite of the remaining days. The mean is of the transformed quantity ln(RR6+1), where RR6 is the rain-rate in mm/6hr – see main text for details.

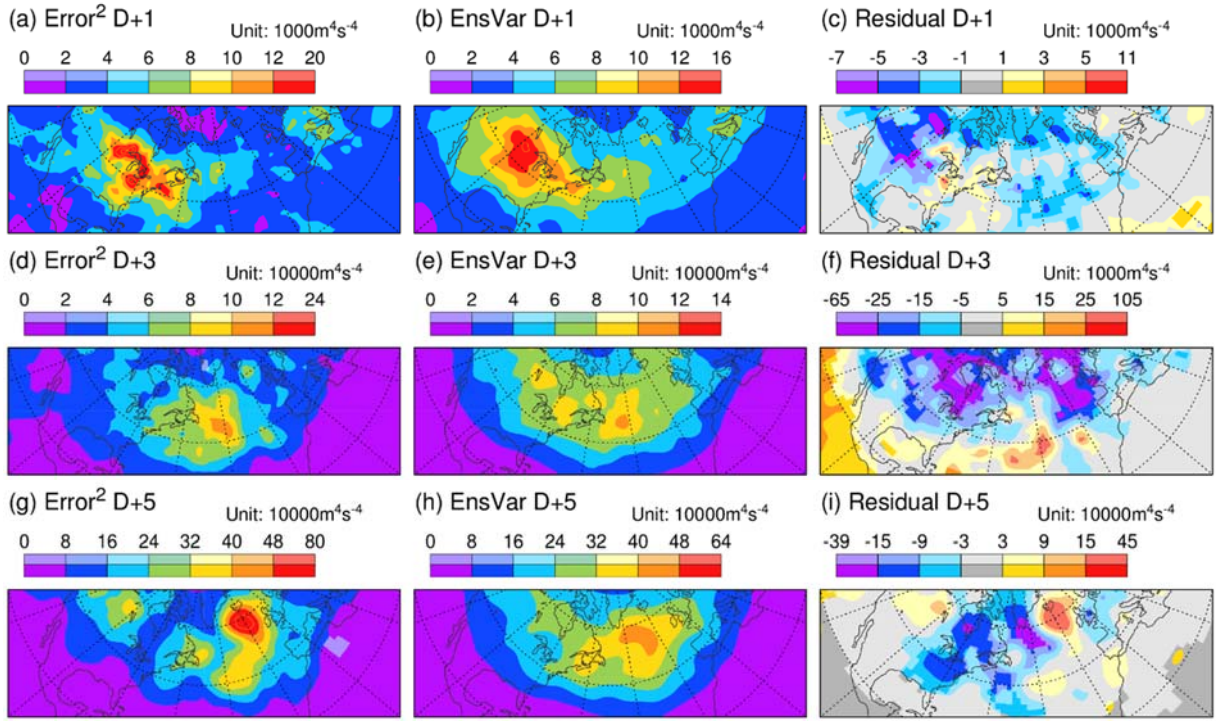

Figure 8: The 'spread-error' relationship (1) within the ENS for the trough/CAPE composite based on 200hPa geopotential. (left) Error^2; the squared-error of the ensemble-mean, (centre) EnsVar; the ensemble variance, scaled for finite ensemble-size and (right) the Residual (=Error^2-EnsVar). D+1 (top), D+3 (middle) and D+5 (bottom). Values significantly different from zero at the 5% level are shown with saturated colours. The unit is indicated in each panel.

The positive residuals over the Great Lakes in Figure 8c suggest that the ensemble variance might be too small in the vicinity of the MCSs, but it is far from convincing that this is the cause of any under-spread to the west of Europe later-on. Regions of negative residuals (particularly D+1 and D+3; Figure

8c,f) are seen in the non-trough/CAPE composite too (not shown), and thus likely to be associated with more general issues.

The crossing of the blue ensemble trajectories in the schematic in Figure 5 represents the fact that errors are growing within a non-linear regime, interacting, and dispersing through the action of teleconnections and waves in general. These effects make it inherently difficult to assess ensemble reliability in the medium-range, and even harder to identify the causes of any unreliability.

To avoid such issues, a new diagnostic based on the EDA has been developed. The much shorter leadtime involved (~12hr) means that errors are growing within a more linear regime, and have not had so much time to interact or to disperse geographically. The hope is that this will reduce the required sample size, and make the assessment of ensemble reliability more local. At these short leadtimes, uncertainty in our knowledge of the truth cannot be neglected. We could incorporate an EDA analysis variance term into our reliability test to take account of this aspect, but we choose to work in observation-space since a good estimation of observation bias and random error represents a more 'foundational' aspect of the data assimilation process. Because observation errors cannot be neglected, we talk about "departures" from observations rather than "errors" from the truth. The new "EDA reliability budget", which is based on the red pentagon in Figure 5, is a decomposition of the mean-squared departures of the form

$$\mathrm{Depar}^2 = \mathrm{Bias}^2 + \mathrm{EnsVar} + \mathrm{ObsUnc}^2 + \mathrm{Residual} \quad . \qquad (2)$$

EnsVar is, as before, the scaled sample variance – but now of the EDA background forecasts. ObsUnc^2 is the sample variance of the observation errors – as assigned within the assimilation system (see, *e.g.*, Hollingsworth and Lönnberg 1986; Desroziers et al. 2005; Bormann and Bauer 2010 for how observation error variances are estimated). Bias^2 is the square of the estimated remaining bias – for example, after the application of variational bias correction (Dee 2004). Note that (2) reduces to (1) if biases and random observation errors are neglected.

A detailed description of the EDA reliability budget is presented in Rodwell et al. (2015). In that paper the authors show that the Residual term in the seasonal-mean EDA reliability budget is able to indicate local deficiencies in reliability. For example, results demonstrate the need for stochastic physics to adequately represent error growth rates in convective regions, but suggest current stochastic physics might be over-active in sub-tropical anticyclone regions where the mid-tropospheric meteorology is largely characterised by time-mean descent and radiative cooling (Held and Hou 1980; Rodwell and Hoskins 1996, 2001; Klocke and Rodwell 2014). The results also demonstrate that increasing the magnitude of surface pressure observation-error over oceans (Ingleby 2010) leads to better diagnosed reliability – a result that might have application in historical re-analysis based on surface pressure observations (Compo et al. 2011), as well as in operational forecasting.

Here we apply, for the first time, the EDA reliability budget in a flow-dependent situation; the trough/CAPE composite. There are two key questions to answer. (a) Is the EDA reliability budget able to identify statistically significant reliability deficiencies when composited on this flow regime? (b) If so, then what are these deficiencies?

First we consider EDA reliability in a more normal situation by computing the budget for the non-trough/CAPE composite. Figure 9 shows the terms in (2) together with the observation density, for 200 hPa (±15 hPa) zonal wind speed measurements made by aircraft, and actively assimilated in (the control member of) the EDA. Aircraft observations are numerous over central North America at this cruising altitude (Figure 9f) and, indeed, they are particularly influential in the data assimilation system. Observation uncertainty (Figure 9d) is computed from the (independent) observation errors assigned within the data assimilation system. When averaged onto a 2^ox2^o grid, the observation uncertainty is naturally smallest where the observation density is largest. The EDA reliability budget decomposes the squared departure term (Depar^2; Figure 9a) into contributions from the bias (Bias^2; Figure 9b), ensemble variance (EnsVar; Figure 9c), observation uncertainty (ObsUnc^2; Figure 9d) and a Residual term (Figure 9e). The spatial structure of the Depar^2 term in the non-trough/CAPE composite largely follows that of the observation uncertainty. There is a more uniform offset that is partly associated with the ensemble variance (and possibly also the bias). However, Figure 9e suggests that there is a small but statistically significant residual term. As discussed by Rodwell et al. (2015), this residual indicates that either the forecast model's error growth-rate is deficient, or that the assigned observation errors are too small.

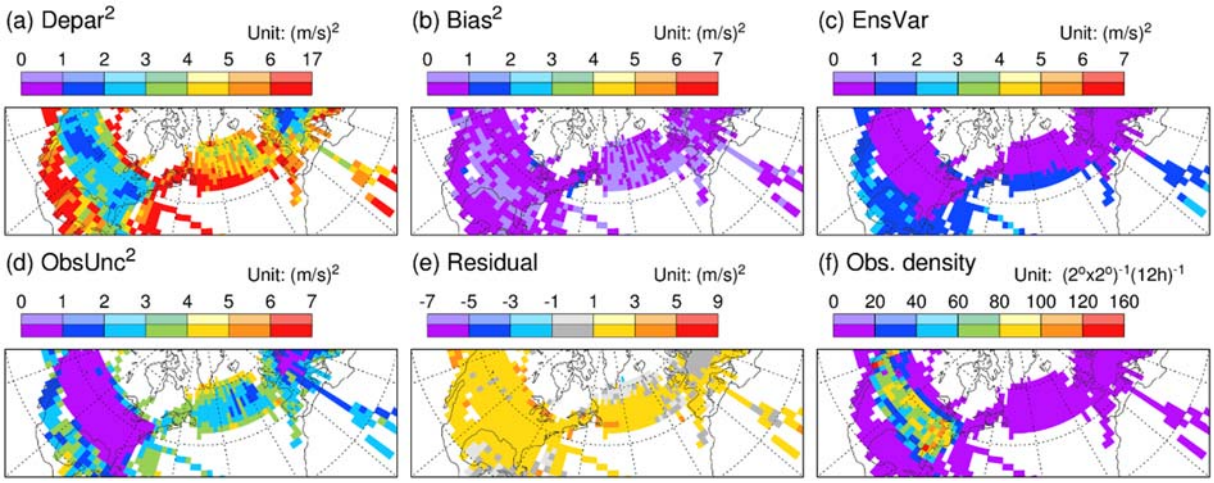

Figure 9: Panels (a) to (e) show the terms in the "EDA reliability budget" for the non-trough/CAPE composite for 200 hPa zonal winds. Data come from aircraft observations between 185 and 215 hPa, and corresponding winds interpolated from the EDA background forecasts. Panel (f) shows the density of aircraft observations assimilated within the EDA control. To reduce noise, an average of at least one observation per (2^ox2^o) grid-box per (12hr) analysis cycle is required for the budget to be plotted. Values significantly different from zero at the 5% level are shown with saturated colours. See main text for more details.

Figure 10 shows the EDA reliability budget terms for the trough/CAPE composite. Comparison of Figure 10 with Figure 9 shows increased departures around the Great Lakes (*c.f.* Figure 10a and 9a) in the trough/CAPE composite. Larger departures are to be expected because of the strong convection taking place. The increased ensemble variance (*c.f.* Figure 10c and 9c) indicates more forecast uncertainty. Notice, however, that this increase is insufficient to fully account for the increased departures, and consequently the Residual term increases markedly too in the MCS region (*c.f.* Figure 10e and 9e). One possibility is that the ObsUnc^2 term does not increase sufficiently in these convective situations. However, aircraft wind observations are thought to be quite accurate (they are assimilated

without bias-correction) and they are probably dense-enough in this region (≥60 per 2°x2° gridbox per 12 hour) to rule-out an increase in representativeness errors for the upper-tropospheric wind-field. Hence the likely conclusion is that the model error growth-rates associated with the MCS/jet-stream interaction are deficient in these trough/CAPE (and MCS) regimes. In other words, model uncertainty is underestimated. The reason(s) for this deficiency remain to be investigated. It could be associated with systematic errors in the height that the convection reaches, but it could also be indicating that, at the resolution of the EDA background model, stronger stochastic physics forcing is required in such convective situations. Note that the EDA background forecasts include the Stochastic Perturbation to Physical Tendencies scheme (SPPT; Buizza et al. 1999, Palmer et al. 2009, Shutts et al. 2011) at the short spatio-temporal scales of interest here, but not the Stochastic Kinetic Energy Backscatter scheme (SKEB; Berner et al. 2009). An interesting sensitivity test would be the inclusion of SKEB– something that is envisioned in the Oxford proposal discussed near the beginning of section 2.

A positive result in terms of diagnostic tool development is that this EDA reliability budget is able to indicate statistically significant flow-dependent reliability deficiencies, and point to their likely causes. Notice, for example, that the residual term in Figure 10e is weak or insignificant outside the MCS region. In the future, it is hoped that the budget will be able to assess the development of more stochastically formulated physical parametrizations, and the effect of increased EDA resolution. Future moves towards developing a fully seamless ensemble analysis/forecast system should hopefully enable deficiencies diagnosed and resolved at short ranges to have a beneficial impact on ensemble reliability at all leadtimes.

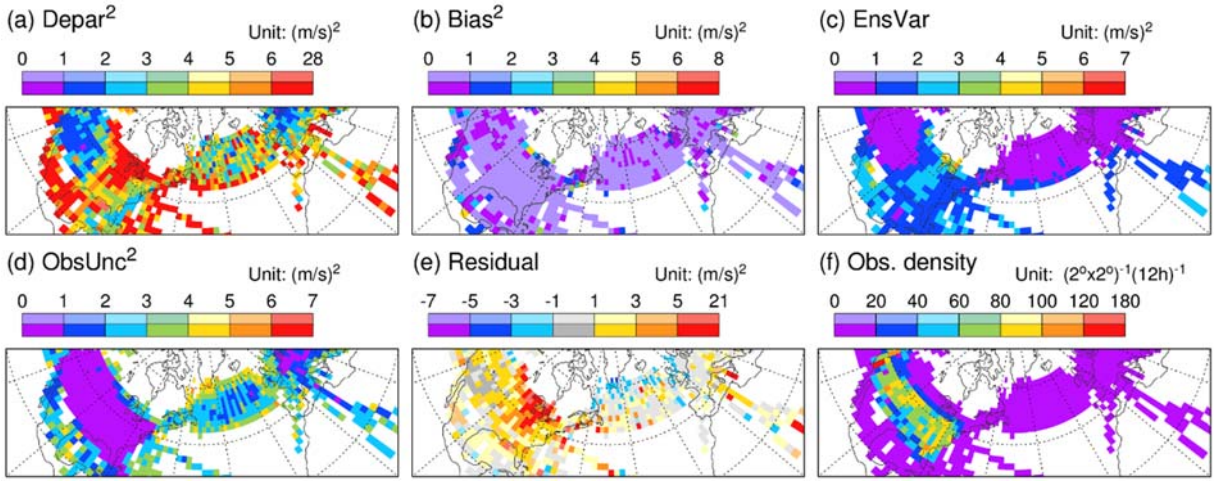

Figure 10: As Figure 9, but for the trough/CAPE composite.

The European forecast busts that sometimes follow MCS events over North America are often associated with a difficult to predict onset of blocking. Blocking is a good example of a "regime" that can persist for several days and can be associated with high-impact cold weather in winter and hot weather in summer. Further diagnosis of regimes and transitions between regimes is given in the next section.

2.3 Diagnostics of regime transitions associated with European cold spells

Heat waves during the warm season and cold-spells during the cold season have a strong societal impact. Over Europe the cold winter of 2009-10 and the warm summer of 2003 caused severe disruption of

activities and heavy loss (or shortening) of life. Droughts, often found in association with heat waves, and wet spells also have a significant economic impact. The prediction of the evolution of such events (onset, maintenance, and decay) a few weeks in advance would be very valuable.

At the sub-seasonal to seasonal time scale, forecasts would not be expected to have skill in predicting day-to-day variability of the weather, but heat-waves /cold-spells, which typically last more than a week, could be the type of events that sub-seasonal prediction systems can predict. Vitart (2005) showed that the ECMWF ensemble (ENS) had some skill in predicting the maintenance of the heat wave during the 2003 summer, but that the predictions of the onset and decay were less successful. It seems that the forecast had a tendency to be excessively persistent. Therefore, it is particularly relevant to investigate the prediction of regime changes leading to such events and their eventual break-down.

It has been noted since early studies (*e.g.* Rex 1950 a,b) that strong and persistent large scale high pressure systems, such as blocking, affect the surface weather in terms of temperature and precipitation. Their occurrences are often associated with dry-spells, and with heat-waves in summer and cold-spells in winter. For example, it was the anti-cyclonic circulation persisting during August 2003 that brought the hot dry tropical continental air mass over Western Europe. While the winter of 2009/2010 was characterized by record persistence of the negative phase of the North Atlantic Oscillation (NAO) which caused several severe cold spells over Northern and Western Europe. The structure of the negative phase of the NAO, has an anticyclonic anomaly over Greenland, a cyclonic anomaly over the Azores, a strong reduction of westerly flow across the Atlantic, and the reinforcement of northerly winds from the Arctic. The large spatial scale and the low-frequency nature of such flow patterns are the key attributes for successful predictions at the sub-seasonal time-scale. In fact, circulation patterns like the NAO are often associated with global teleconnections through propagation of Rossby wave trains. Several studies have shown the lagged correlations between the Madden and Julian Oscillation (MJO) in the tropics and the NAO (Cassou 2008; Lin et al. 2009). Vitart and Molteni (2010) showed that the ENS is able to capture the increase in probability of a positive (negative) NAO following a specific phase of the MJO even beyond 19 days.

Here we explore the ability of the ENS to predict the extended range evolution of the large-scale circulation patterns that are generally associated with cold spells. At present the analysis is limited to the cold months (November to February).

2.3.1 Data and method

We have used coupled ensemble reforecast data from ECMWF (using IFS cycle 40R1; operational in 2014, with the enhanced configuration introduced operationally in 2015), and added a comparison with NCEP (Saha et al. 2014) for part of the study. Salient details are presented in Table 1. In order to have the NCEP ensemble comparable in size with that of ECMWF, we have combined 3 NCEP ensemble forecasts (initiated on consecutive days) into a single 12-member ensemble. (We define the initial date to be that of the central sub-ensemble; this has little effect on results at extended leadtimes).

As verification data, we use ERA-Interim. This choice is not so important at the leadtimes considered here. The variables used are daily fields of geopotential height at 500 hPa and 2m temperature.

Source	ECMWF	NCEP
Period	1995-2014 (20yr)	1999-2010 (12yr)
Frequency	Twice a week	Daily
Ensemble size (members)	11	4
Forecast length (days)	32	60

Table 1: Details of the reforecast data sets used

European wintertime weather is mostly driven by baroclinic instability of the westerly jet stream. As discussed by Cattiaux et al. (2010), the unstable nature of the jet also triggers quasi-stationary circulation patterns of larger scale, often referred to as "weather regimes", which can persist from a few days to a few weeks (Legras and Ghil 1985; Vautard 1990).

The low frequency variability in European temperatures has often been considered as driven by the frequency of occurrence of each regime (Philipp et al. 2007; Vautard and Yiou 2009). For instance the positive (negative) phase of the NAO is generally associated with rather warm (cold) temperatures (e.g., Hurrell 1995), while the persistence of a high pressure system over Northern Europe, often referred to as "Scandinavian blocking " conditions, leads to cold and dry weather over Western Europe (Yiou and Nogaj 2004).

The regimes used in this study have been computed by using the 'k means' clustering algorithm on the distribution of Z500 daily anomalies taken from the ECMWF reanalysis over the domain (80W–40E; 30–90N) for the 29 cold seasons (October–April) 1980–2008 (Ferranti et al. 2014). These regimes are computed as clusters in the phase-space spanned by the ten leading empirical orthogonal functions (EOFs), which explain 80% of the total variance. This EOF pre-filtering removes higher-frequency and smaller-scale variability, and is important to obtain patterns that are reproducible and statistically significant.

Figure 11 shows our four leading regimes. These match well the four well-known Euro-Atlantic regime patterns described in previous studies (Cassou 2008; Dawson et al. 2013). Figure 11(a, b), referred to as +NAO and -NAO, respectively, are consistent with the spatial patterns of the opposite phases of the North Atlantic Oscillation. Figure 11(c), describes anomalous flow during Scandinavian or European blocking events, and so it is referred to as BL. The fourth cluster (Figure 11d) consists of a positive anomaly over the Atlantic Ocean and a negative anomaly over Scandinavia. This pattern is referred to as the Atlantic Ridge (AR).

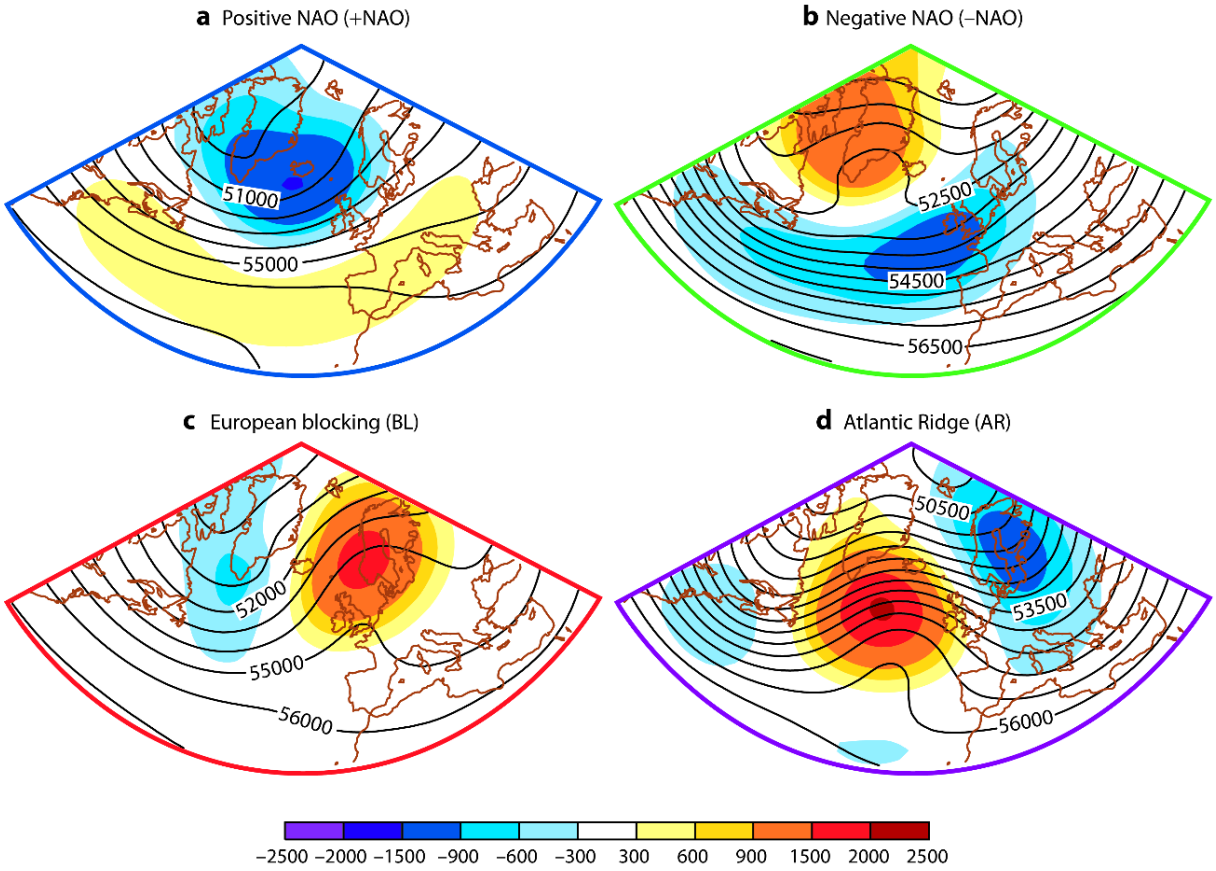

Figure 11: Geographical patterns of the four Euro-Atlantic climatological regimes (both anomalies and full fields). The geopotential anomalies (colour shading) and geopotential (contours) at 500 hPa are shown. Based on cold season (October – April) data for 1980-2008.

Figure 12 shows the 2m temperature anomalies associated with the persistent episodes (> 5days) of each regime for the 20 years of the ECMWF reforecast period. Although each regime is associated with European temperature anomalies, the probability of persisting beyond 12 days for BL and -NAO is about double that of the probability for the other regimes (e.g. Dawson et al. 2012). It follows that BL and -NAO are more likely associated with high-impact temperature anomalies due to their persistence.

In order to study transitions to and from the circulation regimes associated with high-impact temperature anomalies over Europe we explore the use of a 2 dimensional phase space based on the patterns of BL and NAO. Since the regimes patterns in Fig.11 are not strictly orthogonal it is more convenient to use the EOF patterns that arise during our calculation of cluster regimes. Figure 13 shows the leading two EOF patterns. It turns out that ±EOF1 (Figure 13a) resemble quite closely the ±NAO regime patterns (Figure 11ab), the +EOF2 (Figure 13b) resembles the Scandinavian Block (Figure 11c), while the –EOF2 represents, to some extent, the features of the Atlantic Ridge pattern (Figure 11d). In a very simplified way, the leading two EOF patterns represent the symmetric components of the negative/positive NAO and negative/positive BL spatial structure.

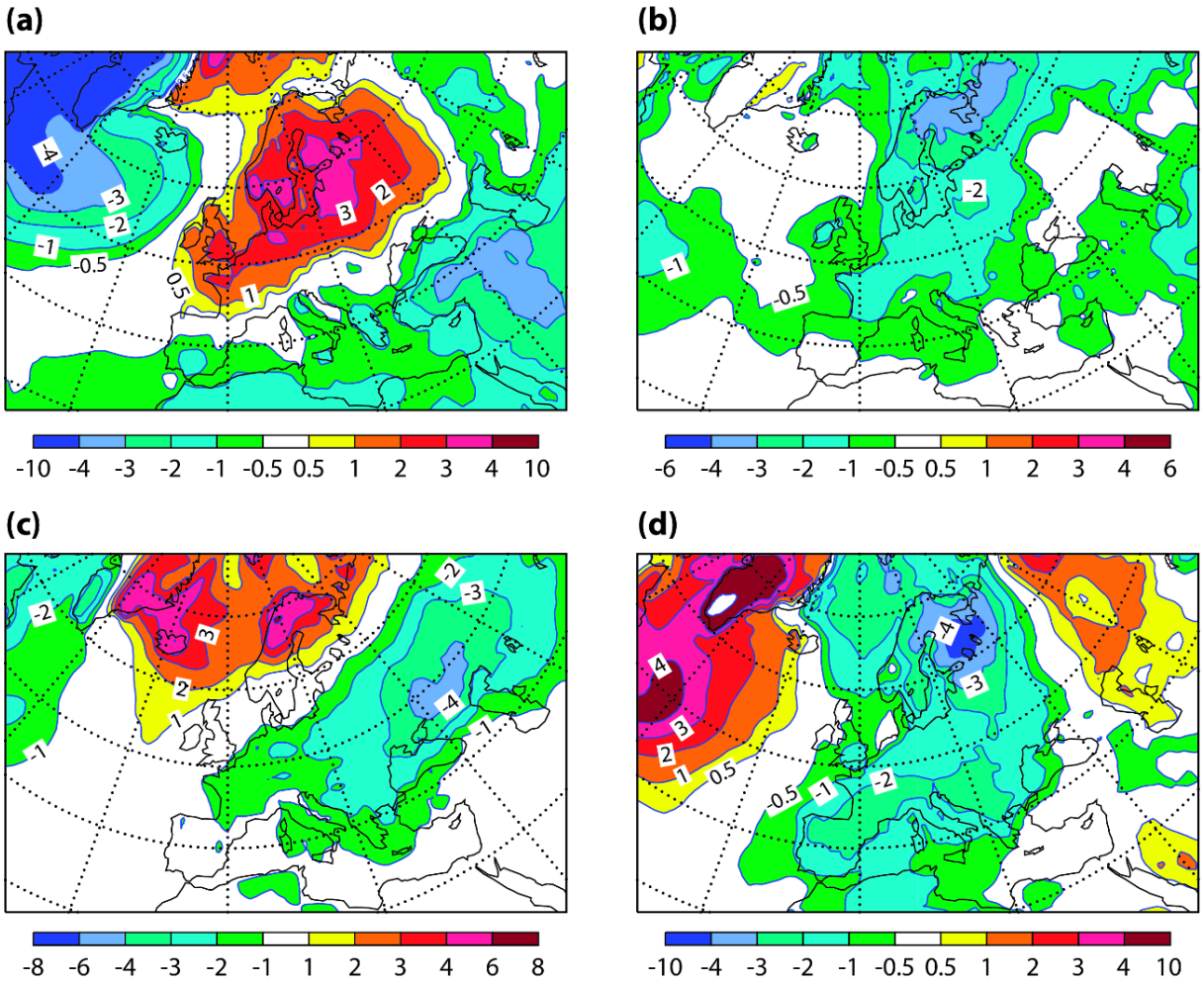

Figure 12: Anomalies in 2m temperature associated with the persistence (periods longer than 5 days) of the (a) positive North Atlantic Oscillation, +NAO, (b) negative North Atlantic Oscillation, -NAO, (c) Scandinavian Blocking, BL, and (d) Atlantic Ridge (AR) regimes.

The two EOFs are used to define a phase space in which the low frequency variability over the Euro-Atlantic region is characterised by the projection onto these two orthogonal patterns of ±NAO and Blocking/anti-blocking (trough over Scandinavia). This very simple view uses the same concept as the well-used MJO index of tropical variability from Wheeler and Hendon (2004). However, while in the tropics two main modes of tropical convection are sufficient to describe the eastward propagation of an MJO event around the globe, the 2-dimensional NAO-BL space can provide just a very partial view of the complex extra-tropical variability. In order to provide an example of the use of the NAO-BL space, Figure 13(c) and (d) show the daily evolution of the analysed geopotential anomalies during the winters 2009-10 and 2013-14, respectively. Figure 13c shows that, during most of December 2009 and February 2010, the flow circulation strongly projected onto the negative phase of the NAO pattern, while in January 2010 the circulation was close to a Scandinavian blocking. Such persistent southward advection of the cold Arctic air resulted in record-breaking cold temperatures over Europe (and is consistent with Figure 12 b and c). In contrast winter 2014 (Fig.13d) was dominated by +NAO and westerly flow anomalies across the Atlantic (Director of Forecasts Report, TAC 46, 2014; ECMWF/TAC/46(14)3).

Consistent with the anomalous flow conditions, winter 2014 was a year of exceptional storminess and severe rainfall, but with rather mild temperatures over Europe.

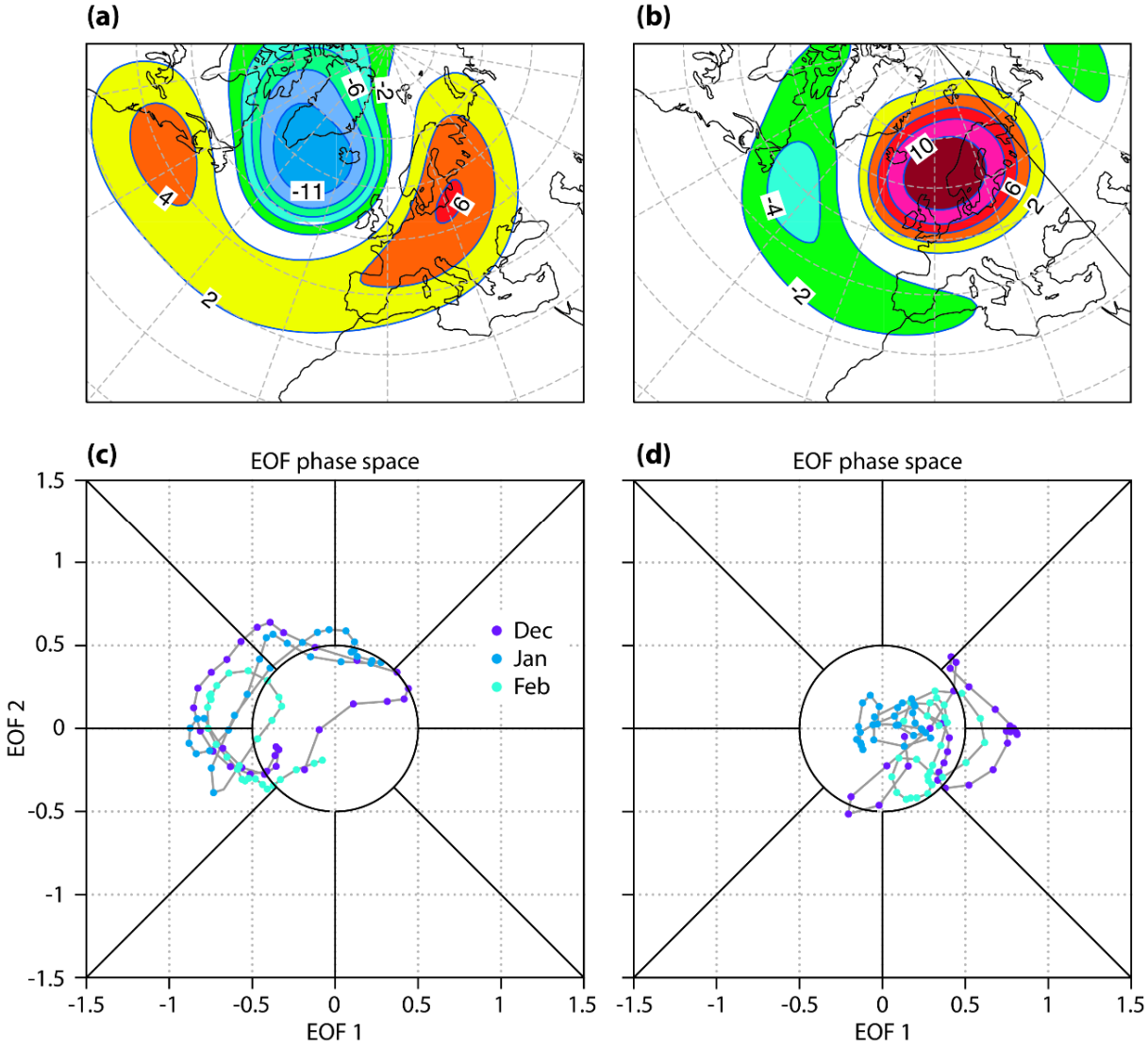

Figure 13: (a) EOF1 and (b) EOF2 of 500 hPa geopotential height, which approximate the +NAO/-NAO and the Scandinavian Block/Atlantic Ridge patterns, respectively. (c,d) Daily evolution of the analysis in the 2-dimensional phase space spanned by the first two EOFs for (c) winter 2009-10 and (d) winter 2013-14.

An important aspect of a model's representation of regimes, that has implications for regime predictability, is the statistics of regime transitions. Despite the large number of studies (Vautard 1990, among others) the problem of finding an optimal definition of regime transitions is still unresolved. In this study we define the transitions based on the values of the projections onto the first two EOFs. We discard transitions between short-lived weather regimes by considering only transitions between regimes that persist for at least 5 days. Figure 14 shows the transition frequencies in/out of blocking (+EOF2, Figure 13b) in ECMWF analysis (filled bars) and re-forecasts (irrespective of leadtime; unfilled bars). The model represents very well the frequency of transitions into blocking from both +NAO (blue) and

-NAO (green) conditions. The model transitions out of the blocking into +NAO (red) and into -NAO (light-blue) suggest a slight over-preference for transitions into the +NAO regime.

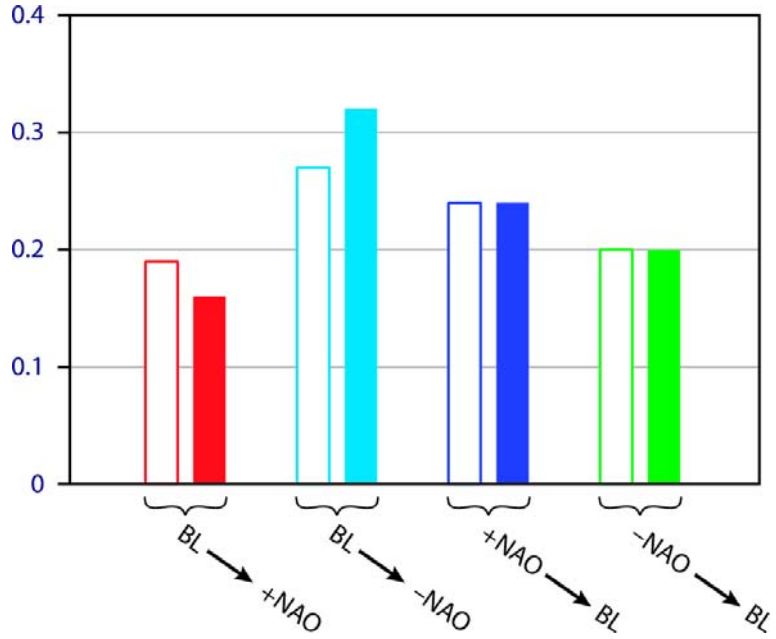

Figure 14: Climatological frequencies of transitions in and out of blocking (BL). Filled bars are for the analysis, unfilled bars are for the model.

2.3.2 Skill comparison in a 2 dimensional (NAO-BL) phase space

The forecast skill in the NAO-BL space is evaluated with the same approach used for the prediction of the MJO index verification. We use anomaly correlation (ACC) and root-mean square error (RMSE) between the observed and ensemble-mean forecast projections onto the first two EOF patterns. The skill as a function of forecast range in predicting ±NAO and ±BL regimes are shown in Figure 15 (a) and (b), respectively, for ECMWF (black) and NCEP (red) systems. The ACC skill in predicting BL, consistent with previous results, drops below 0.5 at about 11 days - a few days earlier than for the NAO. This result is consistent with results of Matsueda and Palmer (2014), and possibly associated with high predictability of the persistence of -NAO regimes. ECMWF has 1-2 days better skill than NCEP for this correlation value. Since beyond day 15 the forecast is not expected to have a day-to-day accuracy, a 5-day running mean has been applied to the projections of both verifying analysis and ensemble mean prior the anomaly correlation computation. The anomaly correlation for the time-filtered values Fig.15 (b) and (d) indicate a smoother decline of the skill as the forecast range increases than in Figure 15 (a) and (c). This is not a surprising result but rather highlights the need for appropriate time/space filtering when we assess the extended ranges (Buizza and Leutbecher 2015).

We have also examined the spread-error relationship for both blocking and NAO projections comparing RMSE and standard-deviation as a function of forecast range (not shown). Beyond the first week both ECMWF and NCEP ensembles are slightly over-confident for blocking and NAO. Although the spread-error relationship is better in the ECMWF system, it is worth mentioning that the re-forecast has a reduced ensemble size in comparison with the real-time forecast (11 vs 51).

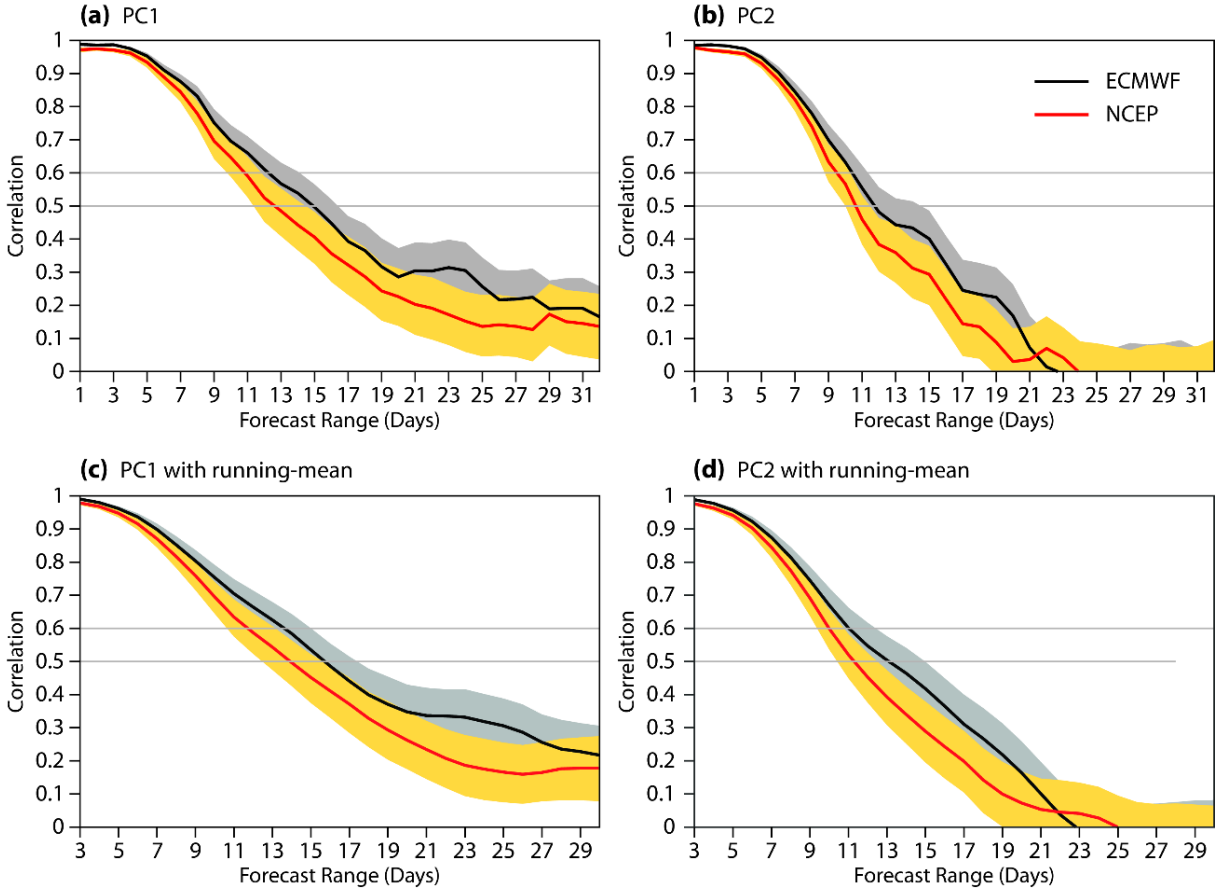

Figure 15: Ensemble mean anomaly correlation as a function of forecast leadtime for the prediction of the principal component associated with (a) ±NAO (westerly/easterly flow across the Atlantic) and (b) Scandinavian blocking/anti-blocking, based on daily ensemble forecasts from ECMWF (black) and NCEP (red). (c,d) are the same as (a,b), but a 5-day running mean has been applied to the forecasts and verifying data. The grey and orange areas represent the 95% level of confidence using a 10,000 bootstrap re-sampling procedure.

This analysis is still on-going. We are planning to characterize better the transitions into blocking by looking at individual cases and composites, including physical process tendencies through our links with ETH Zurich, and investigating the possible association with the tropical flow.

3 Verification of high-impact weather events

An important part of ECMWF's strategy for improving the prediction of high-impact events in the medium and extended range is to create a set of verification metrics suitable for monitoring progress in this regard. Three types of extreme events relevant for Europe are used to illustrate current verification methodologies and outline further developments. These are heavy precipitation events, extra-tropical windstorms, and severe weather associated with organized deep convection. One common theme in the discussion here is the use of additional observation datasets either in the form of higher-density station data, or gridded datasets based on remote sensing observations (radar, lightning detection). In terms of

verification methodology, it is shown how statistical results based on large samples can be related to results obtained from case studies. The scale-dependence of forecast skill can be quantified using the Fractions Skill Score approach (FSS, Roberts and Lean 2008), as a first step towards spatial verification methods. It also provides a framework for future routine verification at ECMWF when the IFS will be run on even higher resolution. This complements the range of standard verification reported in the usual Evaluation paper. New developments not directly related to high-impact weather, such as verification against GPS radio occultation measurements, are reported there.

3.1 Verification of extreme precipitation against high-density observation datasets

ECMWF's routine precipitation verification is based on SYNOP observations, which gives a useful indication of overall precipitation forecast skill (Rodwell et al. 2010; Haiden et al. 2012). For heavy precipitation, however, limitations due to small sample sizes and uneven station distribution become apparent, as well as the representativeness mismatch between measurement scale and model grid scale. Thus ECMWF has started to follow a dual approach by collecting high-density observations, and making increased use of remote sensing data.

3.1.1 Verification using NEXRAD data over North America

The use of radar data for verification of precipitation forecasts at ECMWF has been investigated by Lopez (2014a, b). Comparison of the European ODYSSEY radar composite with SYNOP data shows that ODYSSEY still exhibits substantial biases in some parts of Europe which makes it difficult to use in data assimilation and verification. The NEXRAD Stage IV dataset in the US incorporates both radar and raingauge information. It has been assimilated in the IFS since November 2011, and provides a sufficiently homogeneous and quantitatively reliable precipitation analysis outside the Rocky Mountain region. Because of its high-resolution (4km) and continuous spatial coverage, the up-scaled NEXRAD data allows a better estimation of the grid-scale frequency bias than raingauge observations. A precipitation "event" can be defined by specifying an event "threshold". The event "occurs" if the precipitation exceeds the prescribed threshold. Figure 16(a) shows an evaluation of frequency bias as a function of event threshold. The model predicts the frequency of the event with threshold 20mm well, but for higher thresholds the frequency of occurrence is increasingly underestimated. On the other-hand, there is an over-prediction of light precipitation events, which is commonly known as the 'drizzle problem'. Comparison between the two latest model cycles 40r1 and 41r1 shows the beneficial effect of the cloud microphysics changes in 41r1, leading to a slight reduction of the drizzle problem and a better representation of heavy precipitation (Forbes et al. 2015).

Gridded precipitation analyses permit the evaluation of forecast skill as a function of spatial scale, using for example the Fractions Skill Score (FSS, Roberts and Lean 2008). This is a first step towards the use of spatial verification methods which will become more important for ECMWF as the resolution of the IFS is further increased and "double penalty" effects become stronger in precipitation verification. The FSS measures the squared difference, between a forecast and a gridded observation field, of the fraction of grid points exceeding a given threshold, for a given 'square' area around each grid point (Roberts 2008; Mittermaier and Roberts 2010).

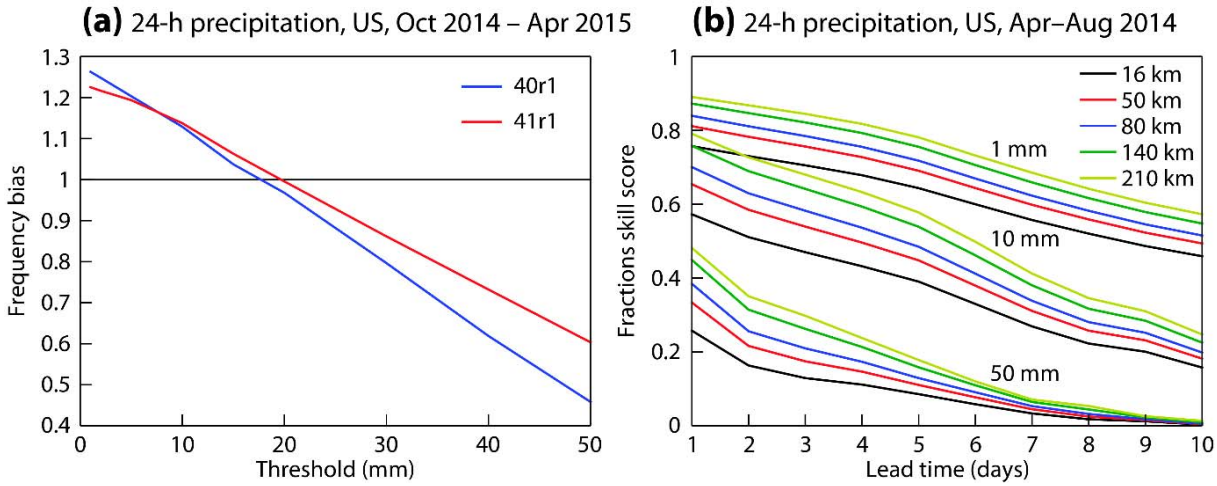

Figure 16: (a) Frequency bias of day-4 HRES forecasts of 24-h precipitation over the US in the period Oct 2014 – Apr 2015 from IFS cycles 40r1 (operational at the time) and 41r1 (operational from 12 May 2015); (b) Fractions skill score for different thresholds and scales during the convective season in 2014.

Figure 16b shows an evaluation of the FSS over the US during the convective season (April – August) 2014 for different thresholds. For 1 mm, which is essentially the distinction between rain and no-rain, the loss in skill with increasing forecast range is moderate, as is the gain in skill with increasing horizontal scale. For a high threshold of 50 mm the strongest drop in skill occurs in the short range. In the short range up to day 3 the curves for 10 and 50 mm thresholds show a larger gain in skill when going from the grid scale (16km, black curves) to larger scales (210km, green curves), than the corresponding gain for the 1 mm threshold. This indicates that for moderate-to-heavy precipitation, location errors are relatively more important than they are for light precipitation.

3.1.2 Verification using high-density observations in Europe

In order to enhance verification of high-impact weather, ECMWF has started to collect high-density observations (HDOBS) from Member States and Co-operating States. Figure 17a shows the geographical distribution of the additional surface observations received by ECMWF as of February 2015. Once the coverage has been further increased, the data will be used to create gridded fields. Even without gridding the HDOBS will help to assess heavy precipitation forecast skill. This is shown using the example of 24-h precipitation observations from the hydrological networks in Europe, which ECMWF receives as part of the European Flood Alert System (EFAS) activities (Pappenberger et al. 2011). Figure 17b shows a domain centred on the south of France, which is one of the Mediterranean areas prone to flood events. It has also been chosen because the EFAS data provides a relatively homogeneous increase in station density in this area compared to SYNOP. Verification results indicate generally lower frequency bias values (not shown), and larger skill for the higher thresholds, when verified against the EFAS data (Figure 18). This is because the EFAS station network tends to be denser in catchments where heavy precipitation occurs more frequently due to orographic effects. In the example given here, these are the parts of the domain located in the Alps and in the Pyrenees. These initial results demonstrate the sensitivity of the verification to the available observational data sets and highlight the importance of comprehensive observation coverage. Further steps will include the merging of available datasets (SYNOP, HDOBS, and EFAS) and an enhanced quality control.

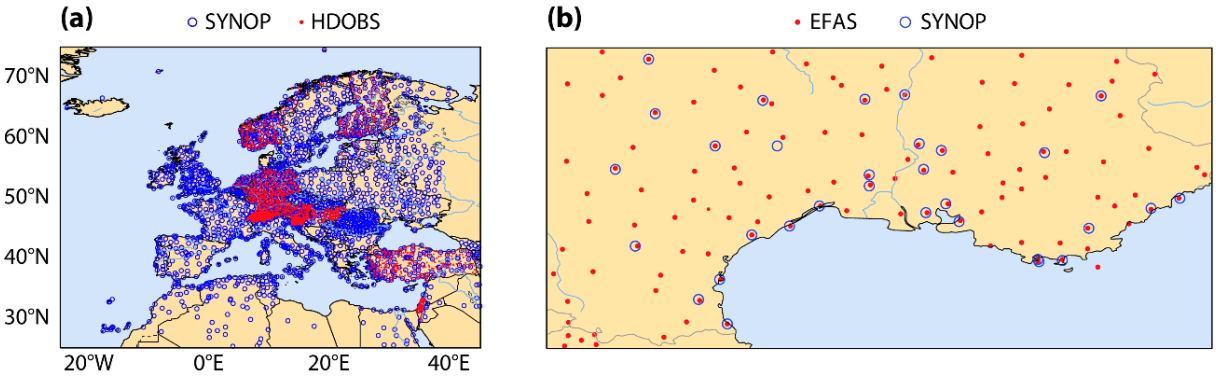

Figure 17: (a) Location of 24-h SYNOP precipitation observations (blue), and high-density observations provided by ECMWF member and cooperating states (red) for February 2015. (b) Location of 24-h SYNOP precipitation observations in southern France (blue) and from the European Flood Alert System (EFAS data, red) for the period June 2013 – March 2015.

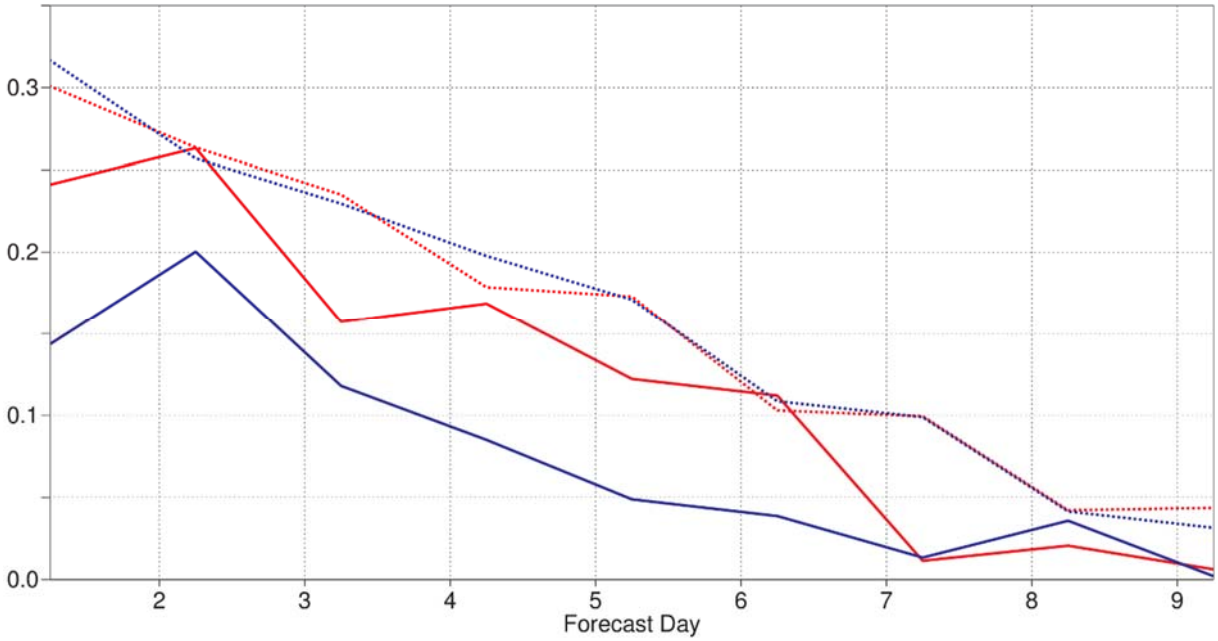

Figure 18: Equitable threat score (ETS) for 24-h precipitation totals above 20 mm (dotted lines) and 60 mm (continuous lines) from 00 UTC runs for the domain shown in Figure 17b in the period June 2013 – March 2015. Verification against SYNOP data in blue, against EFAS data in red.

3.1.3 Verification using high-density observations in Australia

The Australian Bureau of Meteorology (BoM) provides 24-h precipitation observations from raingauges in near real-time for about 1700 stations; about 5 times as many as are available via GTS. The enhanced density of this dataset allows a better assessment of heavy precipitation forecast skill in the region. Here we use it to illustrate the link between single-event skill and overall heavy precipitation skill.

From 20-22 April 2015, New South Wales experienced an exceptionally severe heavy rainfall event associated with an intense, slow moving cyclone. In the most strongly affected area 72-h precipitation totals exceeded 400 mm (Figure 19a). Four casualties were reported. Strong winds and high waves caused additional damage. While the operational HRES gave some indication of a severe rainfall event in the 3-4 day range, the area of maximum precipitation was located off-shore in the forecast (Figure 19b). The new model cycle 41r1, in test mode at that time, (Figure 19c) provided very good guidance in this case, reaching almost 300 mm over land.

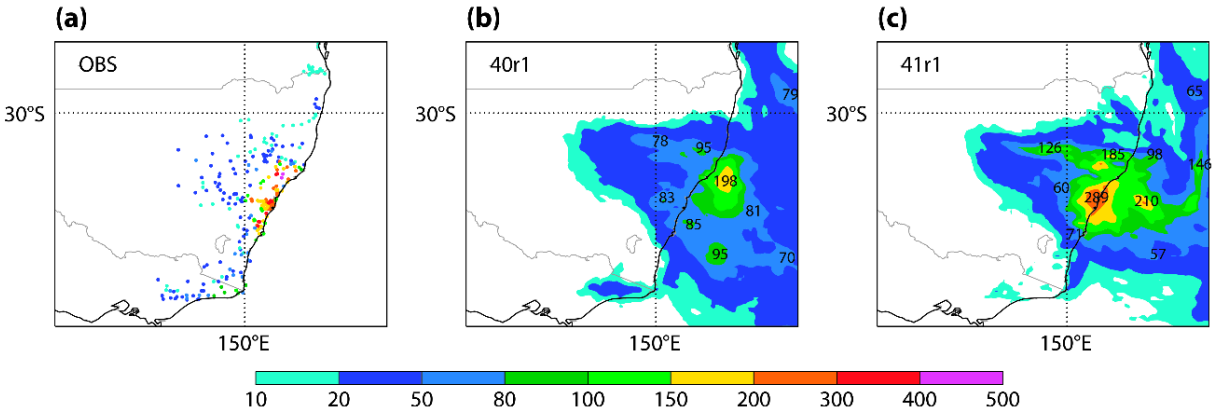

Figure 19: Observed (a) and forecast (HRES) 72-h precipitation totals in south-eastern Australia on 20-22 April 2015 from model cycles 40r1 (b) and 41r1 (c). Forecast range is +24 to +96 h. Precipitation totals in mm.

Since this is a single event, the better forecast from 41r1 could just be fortuitous. We use the SEDI score (Ferro and Stephenson 2011) to link this single event to increasingly larger sets of cases. Figure 20a shows SEDI as a function of lead time for 24-h precipitation totals ≥50 mm for the 3-day period of the event. The improved skill of 41r1 v 40r1 especially on forecast days 2 and 3 is apparent, as well as a rather high skill of 0.8-0.9 on forecast day 1. The curves are noisy because the sample is small but they provide a way of directly comparing the single-event skill with the average skill for similarly intense events. Comparison with results for a period of several months in the same area (Figure 20b) shows that the HRES skill for the event was higher than the period average in the short range, and slightly below the period average in the medium range. Verification for the whole extra-tropics, and for an even longer period (7 months), gives smoother but otherwise surprisingly similar results (Figure 20c). Also, there are relatively small differences between Europe (Figure 20d) and the extra-tropics. This suggests that a SEDI of about 0.7 on day 1, linearly decreasing to about 0.5 around day 8, could be used as a reference, to which individual events can be compared. The comparison also shows that there is a systematic improvement in 41r1 compared to 40r1 but it is smaller than suggested by the single case.

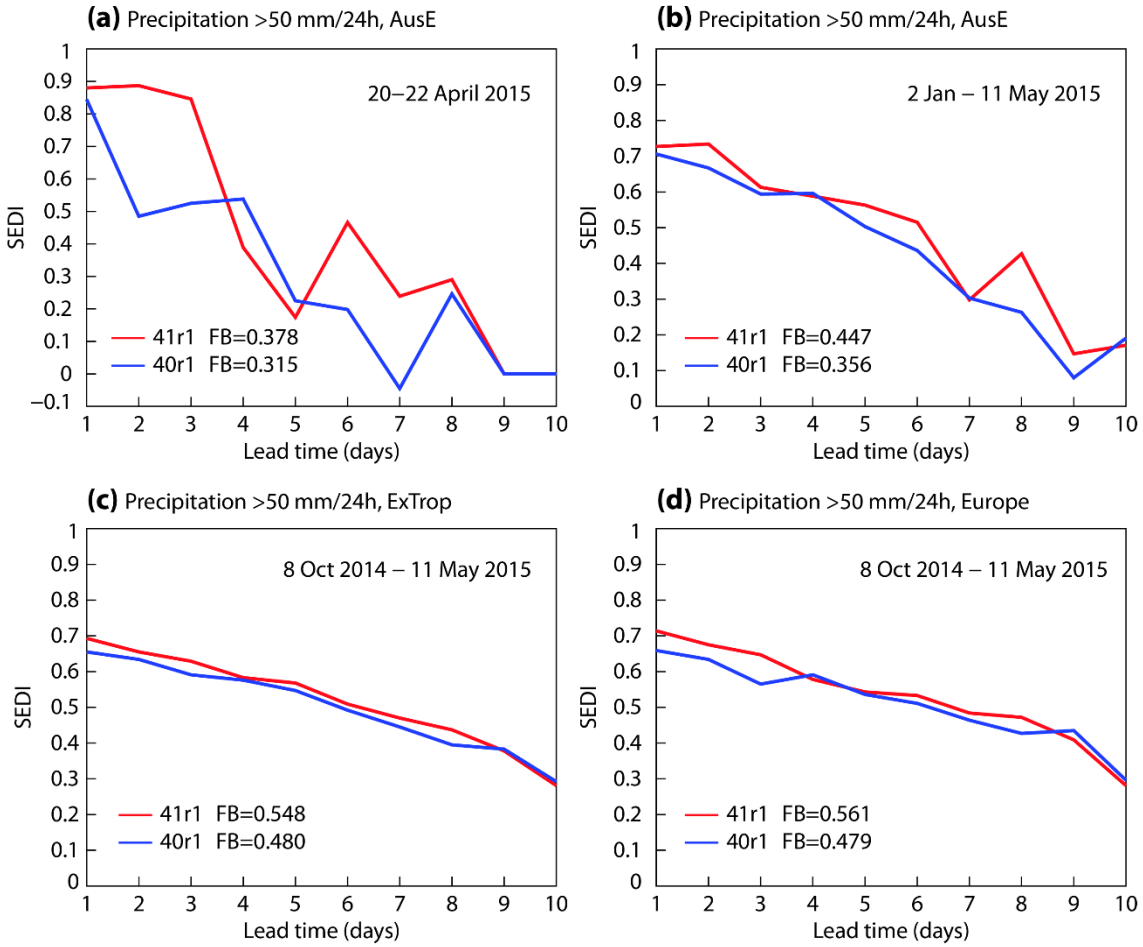

Figure 20: Symmetric extremal dependence index (SEDI) for 24-h precipitation totals >50 mm in eastern Australia (a) during the 3-day period 20-22 April 2015, (b) during the 120-day period 2 Jan – 11 May 2015. Lower two panels show, for comparison, results for (c) the extra-tropics, and (d) Europe. Frequency bias (FB) values averaged over lead times from 1-10 days are given as numbers.

To summarize, the routine verification of extreme precipitation events at ECMWF is currently being enhanced by using additional observation datasets and by creating a methodology which allows us to bridge the gap between case studies and statistics. What has been demonstrated here for the HRES in the forecast range up to day 10 can also be applied (with the appropriate scores, such as BSS or ROC area) to the ENS verification up to day 15, in order to monitor progress towards ECMWF's strategic goals.

3.2 The EFI: condensing and calibrating ENS information in severe convection

Some of the most devastating high-impact weather is associated with severe organized convection. Although the exact timing and location of tornadoes, large hail, or extreme rainfall is difficult to predict in the medium-range, there is considerable skill in the IFS forecasts in delineating areas where there is a high likelihood that severe weather will occur. We demonstrate this by using the Extreme Forecast Index (EFI), applied to the fields of Convective Available Potential Energy (CAPE) and the CAPE-SHEAR parameter (CSP). CSP explores the notion that the likelihood of severe weather and its level of

intensity tend to increase with increasing organization of convection (Tsonevsky 2015). Forecasts are verified against lightning data and severe weather reports.

For Europe, EFI forecasts for CAPE and CAPE-SHEAR have been verified against lightning data from the UK Met Office ATDnet lightning detection system from 1 April to 31 October 2014. Studies suggest that higher lightning activity correlates well with convective severe weather events, and especially with large hail and significant tornadoes (F2 or higher). Figure 23a shows the area under the relative operating characteristic (ROCA) for the EFI for CSP and CAPE at two different intensities of the lightning activity both constituting rare events. The EFI for both parameters, CSP and CAPE, show high skill in discriminating rare and potentially hazardous convective storms. The predictive skill of CSP, which has been designed to specifically identify organized convection, exceeds that of CAPE for the highest-intensity events (continuous blue and red curves), whereas for the less intense events CAPE appears to be the better predictor (dashed blue and red curves). The apparent rise in skill over the first three days for high-intensity events requires further diagnosis.

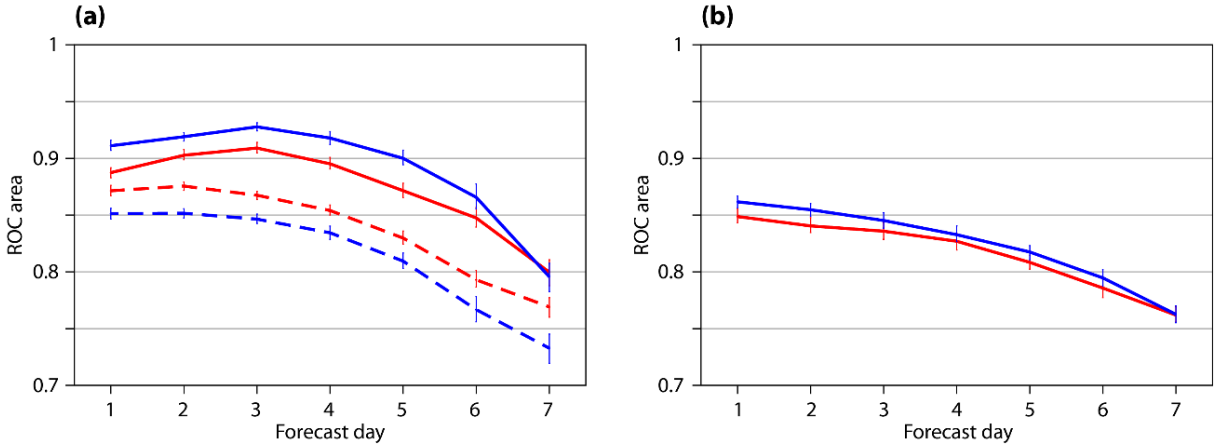

Figure 21: ROC area for the EFI for CSP (blue curves) and CAPE (red curves) over (a) Europe, and (b) the United States. In (a), results for different intensities of lightning activity are shown, with lower-intensity events (dashed curves) corresponding to a frequency of 0.6% in the lightning detection system dataset, and higher-intensity events (solid curves) corresponding to a frequency of 0.04% in the same dataset. In (b), verification against severe weather reports from the Storm Prediction Centre (SPC) is shown. Vertical bars show confidence intervals (5^{th} and 95^{th} percentiles) based on bootstrapping.

A dataset of severe weather reports over the United States has been used for an additional assessment of the ability of the new convective EFI parameters to discriminate between severe and non-severe convection. Reports of tornadoes, large hail (diameter ≥ 2.5 cm) and severe wind gusts (≥ 26 ms^{-1}) for the same mid-year period as for Europe were gridded using the nearest-grid-point method. Figure 23b shows that the EFIs for both CAPE and CSP are able to delineate areas of severe convection in the US at a level of skill comparable to Europe. Consistent with the results for Europe, the EFI for CSP highlights the most severe convective events slightly better than the EFI for CAPE. On average, the values of the EFI for CSP are higher than the EFI for CAPE for severe thunderstorms producing hail at least 5 cm in diameter and/or wind gusts of 33 ms^{-1} or greater.

Verification results indicate that the EFI for CAPE and CSP provides useful guidance on forecasting severe convection in the medium range. ROCA values are well above 0.5 at forecast day 7, which is the

threshold for positive skill. Further work will focus on extending the verification to longer lead times in order to monitor the progress of the IFS in forecasting the potential for severe convection up to two weeks ahead.

The ROC area used here for the verification of severe convection is a special case of the Generalized Discrimination Score D (Mason and Weigel 2009; Weigel and Mason 2011), which is interpretable as an indication of how often forecasts are correct. The D of an unskilled set of forecasts (random guessing or perpetually identical forecasts) is 50%. D can be adapted to various verification contexts, ranging from simple yes–no forecasts of binary outcomes to probabilistic forecasts of continuous variables. Mason and Weigel (2009) have shown that expressions for D in these cases are equivalent to scores that are already known under different names, such as the ROC area, or the Pierce Skill Score. Because of its generality and appealing interpretation, it could be proposed in ECMWF's new strategy as a framework for monitoring progress in forecasting high-impact weather (extreme wind, heat-waves and cold spells) in the extended range.

3.3 The potential economic value of forecasts of extra-tropical windstorms

In the period from October 2013 to March 2014 a number of severe windstorms affected northwest Europe (as discussed in the Director of Forecasts' report to the TAC and SAC last year). Verification of severe wind forecasts at ECMWF included case studies (Hewson et al. 2014) and statistical evaluation (Haiden et al. 2014). The latter study showed that some indication for extreme winds can be provided by the IFS well into the medium range. However, as analysed by Hewson et al. (2014), surface winds can be strongly enhanced by meso-scale features embedded within cyclones, leading to intensity and timing errors in forecasts of maximum wind speeds.

For each forecast parameter there exists a range of possible users with different cost-loss models, and for some of them the detection of extreme events may be of value even if obtained at the cost of a high false alarm rate. This is measured by the potential economic value (PEV), which is also called relative economic value (Richardson 2000). It is computed for cost/loss ratios between 0 and 1, and it shows the economic gain when the forecast is used for decision-making instead of climatology. Figure 21 shows the PEV for Europe in the period October 2013 – March 2014 for (a) moderately strong wind events (80^{th} percentile), and (b) for more extreme wind events (99^{th} percentile), in Europe. For the single forecasts (HRES and CTRL), mitigating action is taken by the user if the event is predicted. For the ensemble forecast (ENS), statistical reliability is assumed, and action is taken if the expected gain is positive. Several features are worth noting:

- Overall, the ENS has the highest PEV, followed by the HRES, and the CTRL.
- The reduction in HRES PEV, when going from the 80^{th} to the 99^{th} percentile, involves both a lower peak value and a narrower range of cost-loss values. For the ENS, it is mainly a narrowing, while the peak values do not change too much, due to the additional degree of freedom offered by the ENS, where each user can choose their optimal probability threshold.
- The relatively high peak PEV values for specific users even at the high percentiles shown here occur despite large false alarm ratios (the fraction of yes forecasts that turn out to be wrong). For example, the PEV value of 0.47 for the 99^{th} percentile from the ENS at day 4 (Figure 21b) is associated with a

false alarm ratio of 0.84. Nevertheless, for users with a small cost/loss ratio of 0.01 these forecasts have substantial value.

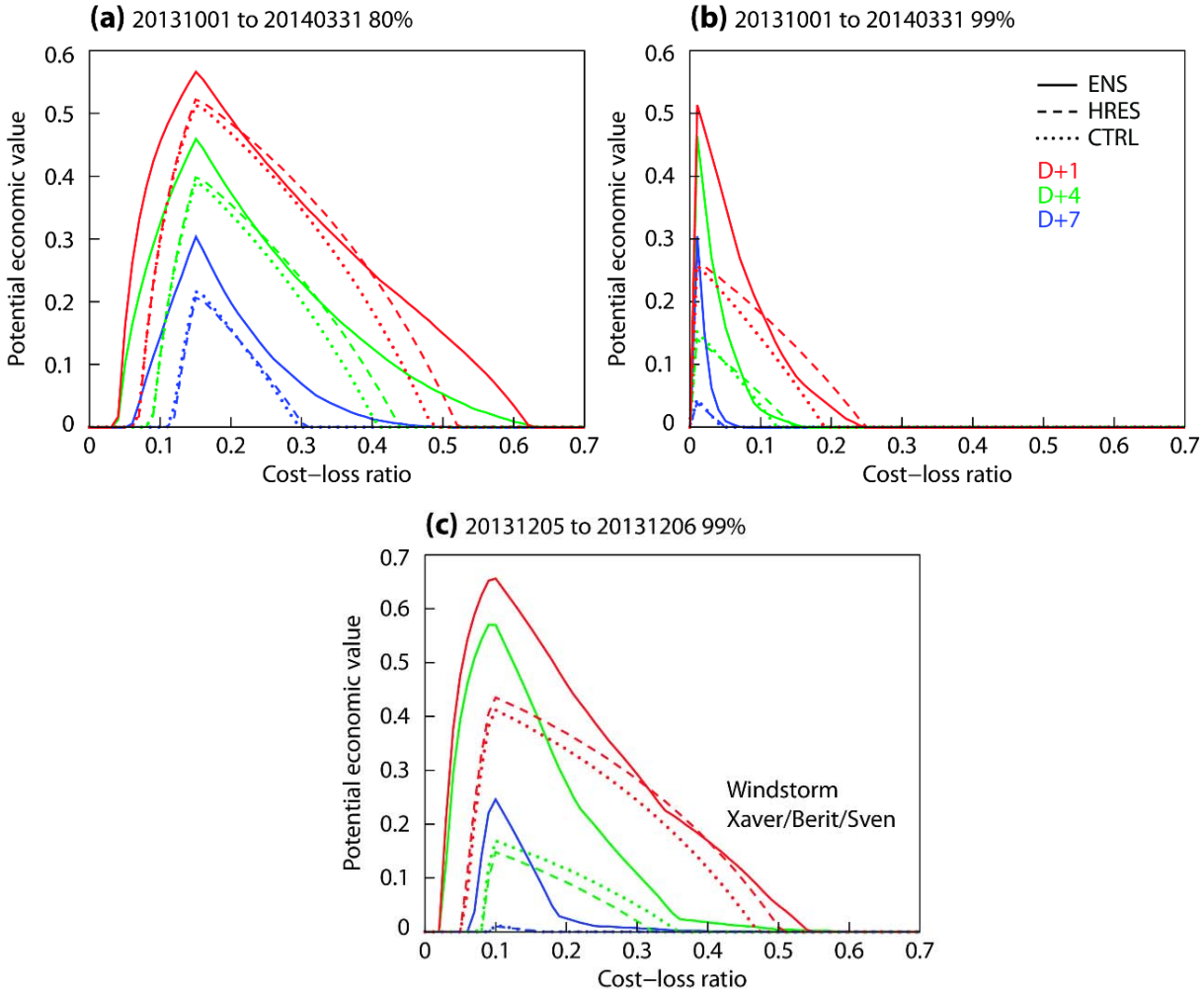

Figure 22: Potential economic value (PEV) of 10-m wind speed forecasts in Europe for events above the (a) 80th and (b) 99th percentile of the climatological distribution. Also shown (c) is the PEV for days on which the observed wind speed at 100 or more stations within the domain exceeded the 99th percentile threshold. The latter reduces the period to the two days 5-6 Dec 2013 of windstorm Xaver/Berit/Sven.

Wind speeds above the 99th percentile (Figure 21b) occur on average once in 100 days (or 3-4 times per year) at each location and grid point. Thus they define a larger, and on the whole a less extreme, set of events than what is usually analysed in case studies, or included in ECMWF's Severe Event Catalogue. Since large-scale windstorms are our main interest here, we set as an additional threshold the number of stations in the domain at which the 99th percentile is exceeded. This number threshold can be increased until the subset of days becomes comparable to the set of events covered in case studies. In this way a link between the statistical evaluation and the case-study methodology can be made.

Figure 21c shows the PEV for the extreme event within the period studied that had the largest-scale impact; windstorm Xaver/Berit/Sven for which over 100 stations exceeded the 99th percentile, and which affected Scotland, Denmark, Germany, and Sweden between 5-6 Dec 2013. The PEV for this event is

actually higher than for the whole period (Figure 21b), at days 1-4 for the HRES and CTRL and days 1-7 (and even day 10, not shown) for the ENS. Notice also that the range of cost-loss ratios with positive skill is also wider. The same results hold for intermediate-scale events (not shown) and this demonstrates that the PEV of forecasts for events that appear in the Severe Event Catalogue is higher than in general.

The sub-sampling of days with large-scale wind events could also be based on the model analysis rather than the SYNOP observations, which would make it potentially more complete. The actual verification should then still be performed against 10-m wind speed observations, since verification against analysis would not reveal the full extent of systematic errors in the model. Figure 22 illustrates this for the large-scale windstorm Xaver/Berit/Sven. The bias at analysis time for wind speeds ≥20 ms^{-1} when verified against SYNOP observations amounts to about –3 ms^{-1} (orographically exposed locations have been excluded).

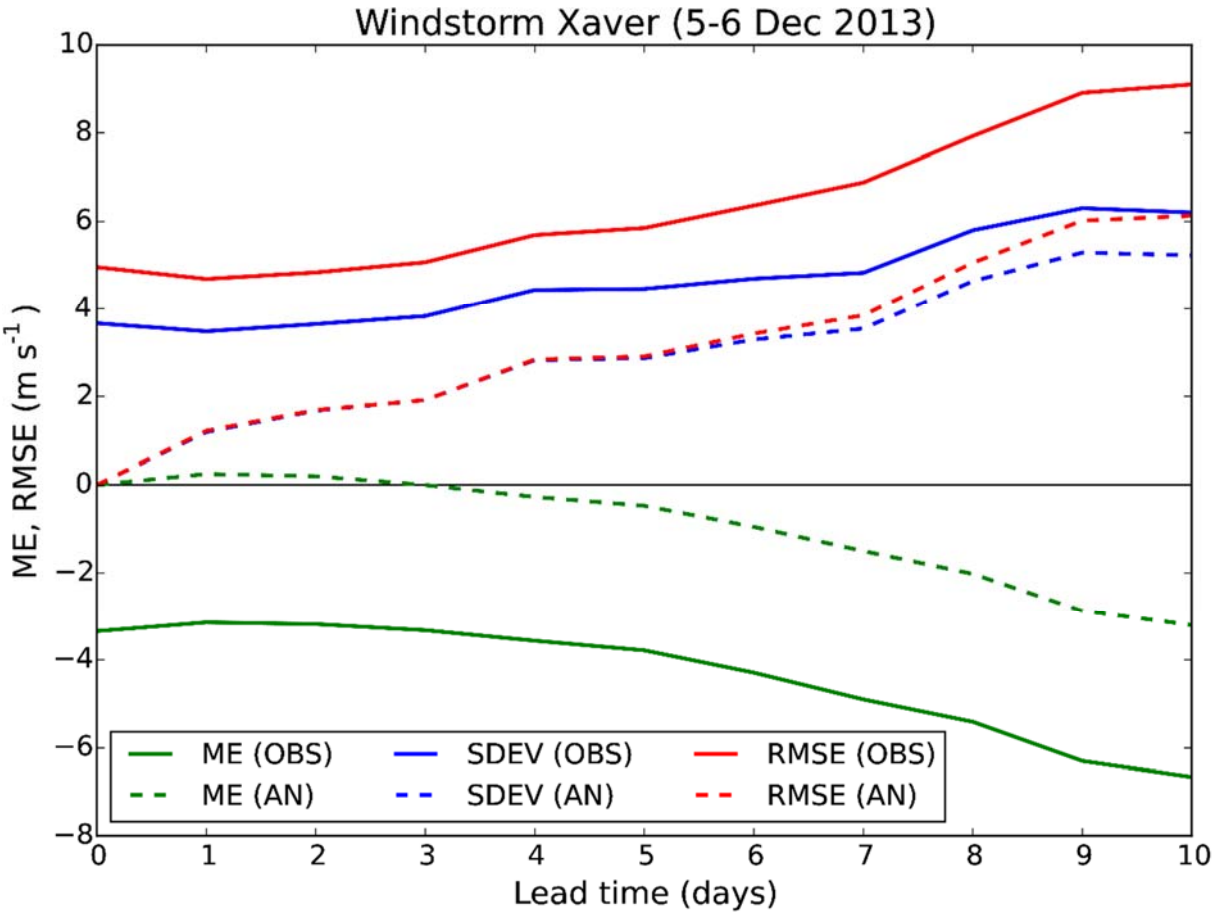

Figure 23: Comparison of HRES forecast errors at 12 UTC computed against SYNOP, and against the model analysis at the same locations. Note that only wind speeds ≥20 ms^{-1} (observed or forecasted) are verified, and only stations at ≤500 m elevation are included.

It has been shown how the PEV of forecasts of individual large-scale windstorm events that occurred contribute positively to the more general statistics for a high (e.g. 99th) percentile. The same analysis could be done based on other scores such as the Brier Skill Score or the ROC area, and the model analysis could be used instead of observations for the identification of cases. The methodology

illustrated here is still based on point-wise verification, however, and measures local skill during large-scale events. As in the case of precipitation, methods such as the Fractions Skill Score approach or spectral filtering (Buizza and Leutbecher 2015) could be applied, in verification against analysis, to measure large-scale forecast skill for extreme events.

4 Conclusions and future directions

To some extent, ECMWF's predictive skill for high-impact weather should increase with more general improvements in the IFS. However, there is a need for an enhanced diagnostic and verification focus on such high-impact weather. This is partly driven by the clear requirement of the user community, and partly because these features pose particular problems throughout the IFS. For example, extreme weather is often associated with small-scale variations that are difficult to resolve, with larger observation errors and observation-representativity errors, with more observation rejections by the data assimilation system, and with extreme physical processes that challenge their parametrization schemes. Finally, high-impact weather involves, almost by definition, small sample sizes, and this requires careful attention. These reasons have been the motivation for much of the work discussed here. With reference to key high-impact weather types (tropical cyclones, extreme extratropical wind and convection, and persistent flow anomalies), we have been able to demonstrate some of the diagnostic tools and verification methods being developed and brought to bear on the IFS.

Diagnostic results have quantified the advantage of using a coupled high-resolution model to represent tropical cyclones, and we have demonstrated a diagnostic tool that can follow such features and produce Lagrangian-mean statistics of (e.g.) background departures. There has been a clear requirement for more diagnostics of the recently introduced ensemble data assimilations (EDA), and this requirement has been met with the development of an EDA reliability diagnostic that assesses our representation of observation uncertainties and model error growth-rates. For extreme mesoscale convective systems, we have quantified deficiencies in error growth-rates that, once addressed, will help improve the flow-dependent reliability of our ensemble forecasts. Better prediction of regimes and regime transitions is high on the users' wish list, and here we have made progress in developing a diagnostic that can characterise the flow-evolution in terms of regime transitions. The utility of this approach will require further investigation. In addition, with the S2S dataset, we can now diagnose and compare the abilities of several operational forecast models to represent and predict regime transitions on the sub-seasonal timescale. ECMWF's progress in delivering its proposed new strategy will require diagnostics to have a strong focus on the ensemble, and to be able to aid the development of a seamless ensemble data assimilation/forecast system. For example there will be further development of diagnostics that can identify extended-range predictability, and diagnostics that can help in the development of more stochastically-formulated physical parametrizations. In order to help deliver skill, diagnostic tools must remain efficient in terms of signal to sample-size, and accurate-enough that results from experimental tests well-predict the eventual impact of IFS changes in the o-suite (operational forecasts). The use of statistical significance testing throughout the diagnostics toolkit is one example of progress towards this end. This toolkit is currently well-applied in the assessment of o-suites and e-suites (experimental forecasts), but there is scope to increase its use at the level of the individual experimenter. Progress has been made in developing the user-interface, and the next step is to strengthen engagement throughout the Centre. Links with the outside community are quite strong, but there are plans to enhance these

further. For example, ECMWF is in the process of engaging an ECMWF Fellow with expertise in the area of diagnostics. There is also the intention to make the diagnostic toolkit available for application to the OpenIFS, which should help the academic community diagnose their work, and provide skill-oriented results that are directly comparable with our operational diagnostics.

Developments in the routine verification of high-impact weather at ECMWF focus on the increased use of non-standard datasets, on the systematic merging of results from single-event case-studies and longer-term statistical results, and on the quantification of the scale-dependence of forecast skill, e.g. using the Fractions Skill Score framework. This methodology will also allow a more unified verification of medium and extended forecast ranges out to seasonal time-scales. The goal is to have a seamless routine evaluation of forecast skill which provides information about skilful time and space scales of various weather parameters at different lead times. This work will include testing of additional scores such as the generalized discrimination score (Mason and Weigel 2009; Weigel and Mason 2011) which, like the potential economic value used here, allows a unified verification of HRES and ENS. This work is also part of ECMWF's activities within the WMO Joint Working Group for Forecast Verification Research (JWGFVR). Another helpful WMO initiative is the extension of the exchange of upper-air scores between global centres to include surface parameters. Proposed thresholds include high values appropriate for high-impact events, such as 15ms^{-1} for 10m wind speed, and 50mm day^{-1} for precipitation. As for upper-air scores, ECMWF as Lead Centre for Deterministic Forecast Verification (LCDNV) will coordinate the exchange.

Acknowledgements

Thanks to Anabel Bowen for help with the figures

References

Aberson, S. D., 2008: Large Forecast Degradations due to Synoptic Surveillance during the 2004 and 2005 Hurricane Seasons. *Mon. Wea. Rev.*, **136**, 3138–3150, doi:10.1175/2007MWR2192.1.

Auligné, T. and A. P. McNally, 2007: Interaction between bias correction and quality control. *Quart. J. R. Meteorol. Soc.*, **133**, 643–653. doi: 10.1002/qj.57.

Baker, N. L. and R. Daley, 2000: Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Quart. J. R. Meteorol. Soc.*, **126**, 1431–1454. doi: 10.1002/qj.49712656511.

Berner J., G.J. Shutts, M. Leutbecher and T.N. Palmer, 2009: A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, **66**: 603–626. doi: 10.1175/2008JAS2677.1.

Bonavita M., L. Isaksen and E. Hólm, 2012: On the use of EDA background error variances in the ECMWF 4D-Var. *Quart. J. R. Meteorol. Soc.*, **138**, 1540–1559. doi: 10.1002/qj.1899.

Bormann N. and P. Bauer, 2010: Estimates of spatial and inter-channel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Quart. J. R. Meteorol. Soc.*, **136**: 1036–1050. doi: 10.1002/qj.616.

Buizza R., M. Miller and T.N. Palmer, 1999: Stochastic representation of model uncertainties in the ECWMF Ensemble Prediction System. *Quart. J. R. Meteorol. Soc.*, **125**, 2887–2908. doi: 10.1002/qj.49712556006.

Buizza, R., and M. Leutbecher, 2015: The forecast skill horizon. ECMWF Technical Memorandum, **754**, 40pp. Available at http://www.ecmwf.int/publications.

Cardinali, C. and R. Buizza, 2004: Observation sensitivity to the analysis and the forecast: A case study during ATreC targeting campaign. In: *Proceedings of the First THORPEX International Science Symposium*, Montreal, Canada.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Quart. J. R. Meteorol. Soc.*, **135**, 239–250. doi: 10.1002/qj.366.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Quart. J. R. Meteorol. Soc.*, **135**, 239–250, doi:10.1002/qj.366.

Cassou, C., 2008: Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation, Nature, **455**(7212), 523–527, doi:10.1038/nature07286.

Cattiaux, J., R. Vautard, C. Cassou, P. Yiou, V. Masson-Delmotte and F. Codron, 2010: Winter 2010 in Europe: A cold extreme in a warming climate, *Geophys. Res. Lett.*, **37**, L20704, doi:10.1029/2010GL044613.

Compo G.P., J.S. Whitaker, P.D. Sardeshmukh, N. Matsui, R.J. Allan, X. Yin, B.E. Gleason, R.S. Vose, G. Rutledge, P. Bessemoulin, S. Brönnimann, M. Brunet, R.I. Crouthamel, A.N. Grant, P.Y. Groisman, P.D. Jones, M. Kruk, A.C. Kruger, G.J. Marshall, M. Maugeri, H.Y. Mok, Ø. Nordli, T.F. Ross, R.M. Trigo, X.L. Wang, S.D. Woodruff and S. J. Worley, 2011: The twentieth century reanalysis project. *Quart. J. R. Meteorol. Soc.*, **137**, 1–28. doi: 10.1002/qj.776.

Dahoui, M., N. Bormann, L. Isaksen, 2014: Automatic checking of observations at ECMWF. *ECMWF Newsletter*, **140**, 21-24. Available at http://www.ecmwf.int/publications.

Davis, C. and co-authors, 2008: Prediction of landfalling Hurricanes with the Advanced Hurricane WRF Model. *Mon. Wea. Rev.*, **136**, 1990–2005, doi: 10.1175/2007MWR2085.1.

Dawson A., T.N. Palmer and S. Corti, 2012: Simulating regime structures in weather and climate prediction models. *Geophys. Res. Lett.*, **39**, L21805, doi: 10.1029/2012GL053284.

Dee D.P., 2004: Variational bias correction of radiance data in the ECMWF system. In: *ECMWF Workshop on Assimilation of High Spectral Resolution Sounders in NWP*. 97–112. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Desroziers G., L. Berre, B. Chapnik and P. Poli, 2005: Diagnosis of observation background and analysis-error statistics in observation space. *Quart. J. R. Meteorol. Soc.*, **131**, 3385–3396. doi: 10.1256/qj.05.108.

Dvorak, V.F., 1975: Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery. Mon. Wea. Rev., 103, 420–430, doi: 10.1175/1520-0493(1975)103<0420:TCIAAF>2.0.CO;2.

ECMWF TAC/46/14: Report by the Director of Forecasts.

Ferranti L., S. Corti and M. Janousek, 2015: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Quart. J. R. Meteorol. Soc.*, **141**, 916 – 924. doi: 10.1002/qj.2411.

Ferro, C.A.T. and D.B. Stephenson, 2011: Extremal Dependence Index: Improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699–713. doi: 10.1175/WAF-D-10-05030.1.

Forbes, R., T. Haiden and L. Magnusson, 2015: Improvements in IFS forecasts of heavy precipitation events. *ECMWF Newsletter*, **144**, 21-26. Available at http://www.ecmwf.int/publications.

Gall, R., J. Franklin, F. Marks, E. N. Rappaport and F. Toepfer, 2013: The hurricane forecast improvement project. *Bull. Amer. Meteor. Soc.*, **94**, 329–343, doi:10.1175/BAMS-D-12-00071.1.

Geer, A. J., P. Bauer and N. Bormann, 2010: Solar biases in microwave imager observation assimilated at ECMWF. *IEE Trans. Geosci. Remote Sens.*, **48**, 2660–2669. doi: 10.1109/TGRS.2010.2040186.

Gopalakrishnan, S., F. Marks, X. Zhang, J.-W. Bao, K.-S. Yeh and R. Altas, 2011: The Experimental HWRF System: A Study on the influence of horizontal resolution on the structure and intensity changes in tropical cyclones using an idealized framework. *Mon. Wea. Rev.*, **139**, 1762–1784, doi: 10.1175/2010MWR3535.1.

Grazzini, F. and L. Isaksen, 2002: North American increments. *ECMWF OD/RD Memo*. 39pp. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Hagedorn, R., R. Buizza, T.M. Hamill, M. Leutbecher and T.N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. R. Meteorol. Soc.*, **138**: 1814–1827. doi: 10.1002/qj.1895.

Haiden, T., L. Magnusson and D. Richardson, 2014: Statistical evaluation of ECMWF extreme wind forecasts. *ECMWF Newsletter*, **139**, 29-33. Available at http://www.ecmwf.int/publications.

Haiden, T., M.J. Rodwell, D. S. Richardson, A. Okagaki, T. Robinson and T. Hewson, 2012: Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Wea. Rev.*, **140**, 2720-2733.

Haiden, T., M. Janousek, P. Bauer, J. Bidlot, L. Ferranti, T. Hewson, F. Prates and D. S. Richardson, 2014: Evolution of ECMWF forecasts, including 2013-2014 upgrades. *ECMWF Technical Memorandum*, **742**. Available at http://www.ecmwf.int/publications.

Halliwell, G.R., S. Gopalakrishnan, F. Marks and D. Willey, 2015: Idealized study of ocean impacts on tropical cyclone intensity forecasts. *Mon. Wea. Rev.*, **143**, 1142–1165, doi:10.1175/MWR-D-14-00022.1.

Hamill T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, doi: 10.1175/1520-0493(2001) 129h0550:IORHFVi2.0.CO;2.

Harnisch, F. and M. Weissmann, 2010: Sensitivity of typhoon forecasts to different subsets of targeted dropsonde observations. *Mon. Wea. Rev.*, **138**, 2664–2680, doi: 10.1175/2010MWR3309.1.

Held I.M. and A.Y. Hou, 1980: Nonlinear axially symmetric circulations in a nearly inviscid atmosphere. *J. Atmos. Sci.*, **37**, 515–533, doi: 10.1175/1520-0469(1980)037h0515:NASCIAi2.0.CO;2.

Hewson, T., L. Magnusson, O. Breivik, F. Prates, I. Tsonevsky and H. J. W. de Vries, 2014: Windstorms in northwest Europe in late 2013. *ECMWF Newsletter*, **139**, 22-28. Available at http://www.ecmwf.int/publications.

Hollingsworth A and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus*, **38A**, 111–136, doi: 10.1111/j.1600-0870.1986.tb00460.x.

Hsiao, L.-F., C.-S. Liou, T.-C. Yeh, Y.-R. Guo, D.-S. Chen, C.-T. Huang, K.- N. Terng, and J.-H. Chen, 2010: A vortex relocation scheme for tropical cyclone initialization in Advanced Research WRF. *Mon. Wea. Rev.*, **138**, 3298–3315, doi:10.1175/2010MWR3275.1.

Hurrell JW. 1995. Decadal trends in the North Atlantic Oscillation, regional temperatures and precipitation. *Science*, **269**, 676–679. doi: 10.1126/science.269.5224.676.

Ingleby B., 2010: Factors affecting ship and buoy data quality: A data assimilation perspective. *J. Atmos. Oceanic Technol.*, **27**, 1476–1489, doi: 10.1175/2010JTECHA1421.1.

Isaksen, L., J. Hasler, R. Buizza and M. Leutbecher, 2010b. The new ensemble of data assimilations. ECMWF Newsletter, **123**, 17—21. Available at http://www.ecmwf.int/publications.

Isaksen L., M. Bonavita, R. Buizza, M. Fisher, J. Hasler, M. Leutbecher and L. Raynaud, 2010a: Ensemble of data assimilations at ECMWF. *ECMWF Technical Memorandum*, **636**. 48pp. Available at http://www.ecmwf.int/publications/.

Ito, K., T. Kuroda and K. Saito, 2015: Forecasting a large number of tropical cyclone intensities around Japan using a high-resolution atmosphere-ocean coupled model. *Wea. Forecasting*, **30**, 793–808, doi: 10.1175/WAF-D-14- 00034.1.

Janssen, P.A.E.M. and co-authors, 2013: Air-sea interaction and surface waves. *ECMWF Technical Memorandum*, **712**. Available at http://www.ecmwf.int/publications.

Jones, S., B. Golding and co-authors, 2014: A research activity on High Impact Weather within the World Weather Research Programme. WMO WWRP, pp87. Available at http://www.wmo.int/pages/prog/arep/wwrp/new/high_impact_weather_project.html

Jones, S.C., P.A. Harr, J. Abraham, L.F. Bosart, P.J. Bowyer, J.L. Evans, D.E. Hanley, B.N. Hanstrum, R.E. Hart, F. Lalaurette, M.R. Sinclair, R.K. Smith and C. Thorncroft, 2003: The extratropical transition of tropical cyclones: Forecast challenges, current understanding, and future directions. *Wea. Forecasting*, **18**, 1052–1092. doi: 10.1175/1520-0434(2003)018<1052:TETOTC>2.0.CO;2.

Jung, T., 2005: Systematic errors of the atmospheric circulation in the ECMWF forecasting system. *Quart. J. R. Meteorol. Soc.*, **131**: 1045–1073. doi: 10.1256/qj.04.93.

Klinker E. and P.D. Sardeshmukh, 1992: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci.*, **49**, 608–627. doi: 10.1175/1520-0469(1992)049<0608:TDOMDI>2.0.CO;2.

Klocke D. and M.J. Rodwell, 2014: A comparison of two numerical weather prediction methods for diagnosing fast-physics errors in climate models. *Quart. J. R. Meteorol. Soc.*, **140**, 517–524, doi: 10.1002/qj.2172.

Knapp, K.R., M.C. Kruk, D.H. Levinson, C.J. Diamond and H. J. Neumann, 2010: The International Best Track Archive for Climate Stewardship (IBTrACS). *Bull. Am. Meteorol. Soc.*, **91**, 363–376, doi: 10.1175/2009BAMS2755.1.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. R. Meteorol. Soc.*, **129**, 3037–3057. doi: 10.1256/qj.02.152.

Landsea, C. W. and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, doi: 10.1175/MWR-D-12-00254.1.

Langland, R. H. and N. L. Baker, 2004: Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus*, **56A**, 189–201. doi: 10.1111/j.1600-0870.2004.00056.x.

Legras, B. and M. Ghil, 1985: Persistent anomalies, blocking and variations in atmospheric predictability, *J. Atmos. Sci.*, **42**, 433–471.

Leutbecher M. and S.T.K. Lang, 2014: On the reliability of ensemble variance in subspaces defined by singular vectors. *Quart. J. R. Meteorol. Soc.*, **140**, 1453–1466, doi: 10.1002/qj.2229.

Leutbecher M. and T.N. Palmer, 2008: Ensemble forecasting. *J. Comp. Phys.*, **227**, 3515–3539. doi: 10.1016/j.jcp.2007.02.014.

Levinson, D.H., H.J. Diamond, K.R. Knapp, M.C. Kruk and E.J. Gibney, 2010: Toward a homogenous global tropical cyclone best-track dataset. *Bull. Am. Meteorol. Soc.*, **91**, 277–280, doi: 10.1175/2010BAMS2930.1.

Lin H., G. Brunet and J. Derome, 2009: An observed connection between the North Atlantic Oscillation and the Madden-Julian oscillation *J. Climate*, 22 364-380. 10.1175/2008JCLI2515.1.

Lopez, P., 2014a: Comparison of ODYSSEY precipitation composites to SYNOP rain gauges and ECMWF model. *ECMWF Technical Memorandum*, **717**, 17pp. Available at http://www.ecmwf.int/publications.

Lopez, P., 2014b: Comparison of NCEP Stage IV precipitation composites with ECMWF model. *ECMWF Technical Memorandum*, **728**, 17pp. Available at http://www.ecmwf.int/publications.

Lorenz E.N., 1963: Deterministic non periodic flow. *J. Atmos. Sci.*, **20**, 130–141. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Lupu, C., C. Cardinali and A.P. McNally, 2015: Adjoint-based forecast sensitivity applied to observation error variances tuning. *Quart. J. R. Meteorol. Soc.* doi: 10.1002/qj.2599.

Magnusson, L., 2015: Model climate diagnostics. *ECMWF seminar proceedings 'Physical processes in present and future large-scale models'*. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Magnusson, L., J.-R. Bidlot, S.T.K. Lang, A. Thorpe, N. Wedi and M. Yamaguchi, 2014: Evaluation of medium-range forecasts for hurricane Sandy. *Mon. Wea. Rev.*, **142**, 1962–1981. doi: 10.1175/MWR-D-13-00228.1.

Martin, J.D. and W.M. Grey, 1993: Tropical cyclone observation and forecasting with and without aircraft reconnaissance. *Wea. Forecasting*, **8**, 519–532, doi: 10.1175/1520-434(1993)008¡0519:TCOAFW¿2.0.CO;2.

Mason, S.J. and A.P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev*., **137**, 331-349. doi: 10.1175/2008MWR2553.1.

Matsueda, M. and T. Palmer, 2014: Predictability of wintertime Euro-Atlantic weather regimes in medium-range forecasts. In *EGU General Assembly Conference Abstracts* (Vol. 16, p. 12945).

Mauritsen, T. and E. Källén, 2004: Blocking prediction in an ensemble forecasting system. *Tellus*, **56A**, 218–228, doi: 10.1111/j.1600-0870.2004.00052.x.

Mittermaier, M. and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343-354. doi: 10.1175/2009WAF2222260.1.

Molteni F. and T.N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. R. Meteorol. Soc.*, **119**, 269–298, doi:10.1002/qj.49711951004.

Molteni, F., R. Buizza, T.N. Palmer and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. R. Meteorol. Soc.*, **122**, 73–119. doi: 10.1002/qj.49712252905.

Palmer, T.N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet and J. Tribbi, 1993: Ensemble prediction. *ECMWF seminar proceedings 'Validation of models over Europe: Vol 1'*. ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK.

Palmer T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. *ECMWF Technical Memorandum*, **598**, 44pp. Available at http://www.ecmwf.int/publications.

Pappenberger F., J. Thielen and M. Del Medico, 2011: The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrol. Processes,* **25,** 1091–1113. doi: 10.1002/hyp.7772.

Philipp, A., P.M. Della-Marta, J. Jacobeit, D.R. Fereday, P.D. Jones, A. Moberg, and H. Wanner, 2007: Long-term variability of daily North Atlantic-European pressure patterns since 1850 classified by simulated annealing clustering, *J. Climate*, **20**, 4065–4095. doi: 10.1175/JCLI4175.1.

Rabier F, H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quart. J. R. Meteorol. Soc.*, **126**, 1143–1170. doi: 10.1002/qj.49712656415.

Rex D.F., 1950a: Blocking action in the middle-troposphere and its effects upon regional climate. I An aerological study of blocking action. *Tellus*, **2**, 196-211. doi: 10.1111/j.2153-3490.1950.tb00331.x.

Rex D.F., 1950b: Blocking action in the middle-troposphere and its effects upon regional climate. II The climatology of blocking action. *Tellus*, **2**, 275-301. doi: 10.1111/j.2153-3490.1950.tb00339.x.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. R. Meteorol. Soc*., **126**, 649–667. doi: 10.1002/qj.49712656313.

Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteorol. Appl*., **15**, 163-169. doi: 10.1002/met.57

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. doi: 10.1175/2007MWR2123.1.

Rodwell M.J. and B.J. Hoskins, 1996: Monsoons and the dynamics of deserts. *Quart. J. R. Meteorol. Soc.*, **122**, 1385–1404. doi: 10.1002/qj.49712253408.

Rodwell M.J. and B.J. Hoskins, 2001: Subtropical anticyclones and summer monsoons. *J. Climate*, **14**, 3192–3211. doi: 10.1175/1520-0442(2001)014<3192:SAASM>2.0.CO;2.

Rodwell M.J. and T.N. Palmer, 2007: Using numerical weather prediction to assess climate models. *Quart. J. R. Meteorol. Soc.*, **133**, 129–146. doi: 10.1002/qj.23.

Rodwell M.J., L. Magnusson, P. Bauer, P. Bechtold, M. Bonavita, C. Cardinali, M. Diamantakis, P. Earnshaw, A. Garcia-Mendez, L. Isaksen, E. Källén, D. Klocke, P. Lopez, T. McNally, A. Persson, F. Prates and N. Wedi, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bull. Amer. Meteor. Soc.*, **94**, 1393–1405, doi:10.1175/BAMS-D-12-00099.1.

Rodwell, M. J., D. S. Richardson, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in NWP. *Quart. J. R. Meteorol. Soc*., **136**, 1344-1363.

Rodwell, M.J. , S.T.K. Lang, N.B. Ingleby, N. Bormann, E. Hólm, F. Rabier, D.S. Richardson and M. Yamaguchi, 2015: Reliability in ensemble data assimilation, *Quart. J. R. Meteorol. Soc.*, Accepted for publication.

Saetra Ø., H. Hersbach, J.R. Bidlot and D. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. Mon. Wea. Rev., 132, 1487–1501. doi: 10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Saha, S. and co-authors, 2014: The NCEP Climate Forecast System Version 2 J. Climate, 27, 2185–2208. doi: http://dx.doi.org/10.1175/JCLI-D-12-00823.1.

Sanders F., 1958: The evaluation of subjective probability forecasts. *Sci. Rept.*, **5**, 63pp. MIT, Dept. of Earth, Atmospheric and Planetary Sciences, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA.

Shutts G., M. Leutbecher, A. Weisheimer, T. Stockdale, L. Isaksen and M. Bonavita, 2011: Representing model uncertainty: Stochastic parametrization at ECMWF. *ECMWF Newsletter*, **129**, 19-24. Available at http://www.ecmwf.int/publications.

Sutton O.G., 1954. The development of meteorology as an exact science. *Quart. J. R. Meteorol. Soc.*, **80**, 328–338. doi: 10.1002/qj.49708034503.

Takaya, Y., F. Vitart, G. Balsamo, M. Balmaseda, M. Leutbecher and F. Molteni, 2010: Implementation of an ocean mixed layer model in IFS. *ECMWF Technical Memorandum*, **622**. Available at http://www.ecmwf.int/publications.

Thompson, P.D., 1957: Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus*, **9**, 275–295. doi: 10.1111/j.2153-3490.1957.tb01885.x

Tsonevsky, I., 2015: New EFI parameters for forecasting severe convection. *ECMWF Newsletter*, **144**, 27-32. Available at http://www.ecmwf.int/publications.

Vautard R., 1990. Multiple weather regimes over the North Atlantic: Analysis of precursors and successors. *Mon. Wea. Rev.*, **118**, 2056–2081. doi: 10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2.

Vautard, R. and P. Yiou, 2009: Control of recent European surface climate change by atmospheric flow. *Geophys. Res. Lett.*, **36**, L22702. doi: 10.1029/2009GL040480.

Velden, C. and co-authors, 2006: The Dvorak Tropical Cyclone Intensity Estimation Technique: A satellite-based method that has endured for over 30 Years. *Bull. Amer. Meteor. Soc.*, **87**, 1195–1210, doi: 10.1175/BAMS-87-9-1195.

Vitart F., 2005: Monthly forecast and the summer 2003 heat wave over Europe: a case study. *Atmos. Sci. Lett.*, **6**, 112-117. doi: 10.1002/asl.99.

Vitart F. and F. Molteni, 2010: Simulation of the Madden and Julian oscillation and its teleconnections in the ECMWF forecast system. *Quart. J. R. Meteorol. Soc.*, **136**, 842-855. doi: 10.1002/qj.623.

Vitart, F., F. Prates, A. Bonet and C. Sahin, 2012: New tropical cyclone products on the web. *ECMWF Newsletter*, **131**. Available at http://www.ecmwf.int/publications.

Wedi, N.P., P. Bauer, W. Deconinck, M. Diamantakis, M. Hamrud, C. Kuehnlein, S. Malardel, K. Mogensen, G. Mozdzynski and P.K. Smolarkiewicz, 2015: The modelling infrastructure of the Integrated Forecasting System (IFS): Recent advances & future challenges. *ECMWF Technical Memorandum*, **760**, in preparation.

Weeler M. and H.H. Hendon 2004: An all-season real-time multivariate MJO index: development of an index for monitoring and prediction. Mon. Wea. Rev., 132, 1917-1932. doi: 10.1175/1520-0493(2004)132<1917:AARMMI>2.0.CO;2.

Weigel, A.P. and S.J. Mason, 2011: A generalized discrimination score for ensemble forecasts. *Mon. Wea. Rev.*, **139**, 3069-3074. doi: 10.1175/MWR-D-10-05069.1.

WMO, 2013: Verification methods for tropical cyclone forecasts. WWRP 2013-7, WMO.

Yablonsky, R.M. and I. Ginis, 2009: Limitation of one-dimensional ocean models for coupled hurricane-ocean model forecasts. *Mon. Wea. Rev.*, **139**, 4410–4419, doi: 10.1175/2009MWR2863.1.

Yiou, P. and M. Nogaj, 2004: Extreme climatic events and weather regimes over the North Atlantic: When and where?, *Geophys. Res. Lett.*, **31**, L07202, doi: 10.1029/2003GL019119.

Zuo, H., M.A. Balmaseda and K. Mogensen, 2015: The new eddy permitting ORAP5 ocean reanalysis: description, evaluation and uncertainties in climate signals. *Clim. Dyn.*, Accepted, doi: 10.1007/s00382-015-2675-1.5.