TECHNICAL MEMORANDUM

796

Diagnostic methods for understanding the origin of forecast errors

Linus Magnusson

Forecast Department

February 2017

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

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, RG2 9AX, England

Abstract

Although the quality of medium-range forecasts has increased considerably over the decades since the start of operational forecasts at ECMWF, individual forecasts still occasionally experience very large errors. Often the phrasing 'drop-outs' or 'forecast busts' is used for such episodes. The aim of this report is to use a combination of methods to track errors in three cases of extreme forecast errors between 2014 and 2016 to better understand the error sources. Manual error tracking and ensemble sensitivity are used to give a first guess of the source region and relaxation experiments are used to confirm the result. In the three investigated cases the errors originated from the tropical eastern Pacific, western/central Canada and western Atlantic respectively. The mechanisms behind the errors are discussed in the report. The results from this study can form a basis for further investigations of these cases and the methodology explained can be applied to understand future bust cases to increase our knowledge on origin and propagation of forecast errors.

1 Introduction

Although the quality of medium-range forecasts has increased considerably over the decades since the start of operational forecasting at ECMWF primarily through improvements in the model formulation and data assimilation (Magnusson and Källén, 2013; Bauer *et al.*, 2015), individual forecasts still occasionally experience very large errors. Often the phrasing 'drop-outs' or 'forecast busts' is used for such episodes. As these episodes contribute negatively to the average quality measures of the forecasting system and are often disturbing for the forecast users, it is meaningful to understand the nature of these episodes. One example of an investigation triggered by a bust is documented in Rodwell *et al.* (2013). The aim of this report is to use a combination of methods to track errors in three cases of extreme forecast errors between 2014 and 2016 to better understand the error sources. The cases can form a basis for further predictability studies.

In this report we focus on large-scale errors over Europe in the medium-range. Hence, we use the errors in 500 hPa geopotential height 6 days into the forecast and averaged over Europe. For this metric the largest errors occur in autumn, winter and spring-time. As the errors often have a large-scale impact on the temperatures, they often have widespread impact on applications such as the energy market and can also deteriorate forecasts for extreme weather (see e.g Lamberson *et al.* (2016)). The climatology of European forecast busts has recently been discussed in Lillo and Parsons (2016).

In order to locate the origin of the errors, different error tracking techniques are available. The most straightforward method is to plot the error (difference between forecast and analysis) for different lead-times and track the error manually backwards in time. In a similar way one can plot the difference between consecutive forecasts to see the propagation of the difference. The latter method will not be discussed in this report but has been used in the investigation of the cases together with the other methods. It is well known that the forecast errors can propagate with the group speed of the Rossby waves and is not bound to the phase speed of the waves (see e.g Kelly *et al.* (2007)), which can lead to impact over Europe from for example initial state errors in the Pacific in 6-day forecasts.

As ensemble forecasts have become much more reliable and better able to predict the evolution of the forecast uncertainties in the atmosphere, ensemble sensitivity methods have become useful to link sensitive structures between different lead times. The ensemble sensitivity could be applied by calculating the covariance (or correlation) between a response function and the forecast perturbation fields that do not have to be the same as used for the metric (as in e.g. Torn and Hakim (2008); Zheng *et al.* (2013)). Similarily, one can rank the performance of each ensemble member using the response function and make composites (as in e.g. Torn *et al.* (2015); Lamberson *et al.* (2016)). In this study we will use both

methods and apply the forecast error over Europe as the response function. The response function does not necessarily have to be related to the error but can for example be based on other forecast properties such as the leading EOFs of the ensemble spread as in Zheng *et al.* (2013).

The third method to be applied is a relaxation (nudging) technique. The concept is to nudge the forecast towards the true state inside a region, and evaluate the downstream influence in the forecast. The technique has been applied in seasonal and monthly forecasts (Jung *et al.*, 2010a,b), mainly to investigate teleconnections from the tropics to the extra-tropics. The techinque was also used in Jung *et al.* (2014) to determine the influence from the Arctic to mid-latitudes in the medium-range. Jung (2011) used the relaxation technique to create initial conditions with the information of the 'truth' limited to a region, and the result proved that the relaxation technique was able to simulate the influence from the region on downstream forecast skill. The relaxation method has recently been used in Lamberson *et al.* (2016) to investigate a case of a falsely predicting a severe windstorm over western Europe.

In this report we combine the three techniques to find source regions for large medium-range errors. The manual error tracking and the ensemble sensitivity will be used to give a first guess of the region to use for relaxation experiments. The methods will be applied for three cases of extreme errors over Europe. The cases selected here are among the worst busts in the period 2014-2016 for ECMWF high-resolution forecast (HRES). In this report we will use HRES together with the 50 member ensemble forecast (ENS). During the period HRES had a spectral resolution of T_l1279 corresponding to 16 km and ENS had resolution of T_L639 corresponding to 32 km. The structure of the report is as follows: In Section 2 the ensemble sensitivity method and relaxation technique are introduced. In Section 3 the charactaristics of the error, the result from the error tracking methods and conditions in the source region for the error are presented for each case. Finally, the results are discussed in Section 4.

2 Methods

2.1 Ensemble sensitivity

The idea behind ensemble sensitivity methods is to use the ensemble to find flow-dependent geographical connections between different lead-times. It can be done by (1) calculating the covariance between a response function J and grid-point values $F(x,y)$ for a different time-step or (2) clustering the ensemble members according to their ranking using the response function. In this report we define the response function as the RMSE in 500 geopotential height (z500) for 6-day forecasts over Europe. The metric is calculated for each ensemble member.

By normalising the covariance with the standard deviation of the ensemble in each grid point and the response function, the correlation is obtained. The correlation is calculated between the response function J and the forecast field F of interest:

$$Corr(x,y) = \frac{\sum_i (J_i - J_{mean})(F_i(x,y) - F_{mean}(x,y))}{\sigma(J)\sigma(F_{all}(x,y)} \quad (1)$$

.

where i is the index of ensemble members and x and y are grid-point coordinates. Note that the forecast field F does not need to be the same parameter and time-step as used for the response function J.

As well as calculating the covariance (or correlations), ensemble sensitivities can also be calculated by comparing clusters of ensemble members. The clusters can for example be defined from the ranking of

the error for each ensemble member. The rank method is here the normalised difference between the mean of the 5 members with the highest rank (error) and the 5 with lowest rank:

$$Diff(x,y) = \frac{(F_{high}(x,y) - F_{low}(x,y))}{\sigma(F_{all}(x,y)} \tag{2}$$

.

where $F_{high}(x,y)$ is the mean of the forecasts in grid-point (x,y) with highest rank and $F_{low}(x,y)$ the mean of the forecast with the lowest rank of the metric.

In this report we compare the correlation method, similar to the one used in Zheng *et al.* (2013) and the rank method used in Torn *et al.* (2015) and Lamberson *et al.* (2016). The relation between ensemble sensitivities (in the covariance form) and adjoint sensitivities are discussed in Ancell and Hakim (2007) and the method was used in Torn and Hakim (2008) to evaluate sensitivity to observations. We are calculating the sensitivities from the response function (RMSE for z500 over Europe Day 6) to z200 (z500 for Case 3) both to forecast-day 6 to see how the sensitivity method reproduces the final error pattern and forecast-day 2 in order to find the sensitive region for forecast errors in short-range forecasts.

2.2 Relaxation technique

Relaxation experiments nudge the model forecast toward the 'truth' (e.g., from an analysis) during the course of the model integration. By constraining the forecast inside a regional box to the true outcome, one can evaluate the connections to other parts of the globe. With a perfect forecast inside the box one expects the error to be reduced downstream, altought it is not guaranteed that the error originates from the box. If the error originates upstream but during the integration propagates into the box, the forecast will be corrected. Therefore is it not needed to restrict the box extention upstream of the suspected region of the initial error. Note that the impact of the relaxation is dependent on the quality of the analysis. As the analyses inside the suspected regions probably suffer from errors close to the initialisation of poor forecasts, the relaxation box needs to be extended downstream of the suspected region.

Relaxation involves adding an extra term of the following form to the prognostic equations of the model:

$$-\lambda(x - x_{ref}) \tag{3}$$

.

where x is the ensemble state vector and x_{ref} is the reference vector toward which the model is drawn (here the analysis interpolated in time). In addition, λ controls the strength of the relaxation and has units of (time step)$^{-1}$. For the relaxation experiments employed in this study, $\lambda = 1/3$, indicating that at each time step, the model is corrected using 33.33% of the departure of x from x_{ref} for all variables and levels. The time-step of the experiments is set to 20 minutes. The strength of the relaxation is smoothed around the edges of the box.

To apply the technique one first has to define the relaxation region. Here the result from ensemble sensitivity and the manual error tracking is useful to get a first guess of the propagation of the error. The relaxation has been applied to ensemble forecast experiments with 20 members, using ECMWF model version 43r1 for Case 1 and 42r1 for Cases 2 and 3. The model uses a horizontal resolution of T_L639. No model perturbations have been applied to the experiments. The relaxation experimentation (in the ECMWF configuration) uses the same vertical configuration as the HRES analysis to avoid vertical interpolation. As the vertical resolution is not the same as for the operational ensemble as the HRES

analysis, an experiment without relaxation has been undertaken as well. For each case several relaxation boxes have been tested, but in this report we only compare the forecasts from the largest box without any significant (small) impact with the smallest box with large impact on the 6-day error over Europe.

3 Results

3.1 Case 1 - 15 March 2014 00UTC

The first case presented here is from March 2014. Figure 1(a-b) shows a 2-month time-series centred on the date for the bust case of root-mean-square-error (RMSE, a) and anomaly correlation coefficient (ACC, b). The scores are calculated for 500 hPa geopotential height (z500), for 6-day forecasts over Europe (35°N-75°N, 12.5°W-42.5°E). The figures include high-resolution forecasts (HRES, red line), ensemble control forecast (black line), ensemble mean (blue, dashed line) and the distribution of errors in individual ensemble members (grey: all members, yellow: 25th-75th percentile). The mean values over the plotted 2-month period are for ACC 86% and RMSE 72 metres, but for the forecast from 15 March 00UTC the scores were -20% for ACC and 214 metres for RMSE. These numbers can be compared to the definition of European busts used in Rodwell *et al.* (2013) where the threshold was ACC below 40% together with RMSE above 60 metres. This bust affected relatively few subsequent forecasts. The performance of HRES was similar to the worst ensemble members, while several ensemble members showed relatively small errors, and the ensemble mean had an ACC of 55%.

Figure 1(c) shows forecasts from different initial times for the daily mean 2-metre temperature in a box centred over Germany (45°N-55°N, 5°E-15°E) valid on 21 March (the verification date of the 6-day forecast from 15 March), for the analysis (green dot), ensemble median (black dot) and probability distribution (blue box-and-whisker) and the HRES forecast (red dot). The figure also includes the probability distribution from the model climate calculated from a 20-year reforecast dataset (red box-and-whisker) and the climate median (dashed line). The choice of Germany is somewhat arbitrary, and it could well be that the forecast bust had a larger impact elsewhere. For the earliest ensemble forecast (from 15 days before the event), the ensemble distribution is similar to the model climate distribution, indicating no predictable signal at this forecast range. Around 9 days before the event (13 March), ensemble forecasts started to indicate a warm anomaly but still with some members being colder than normal. The HRES forecast (red dot) was among the warmest ensemble members but on 15 March the HRES appeared to be among the coldest members, which turned out to be the erroneous solution and resulted in the large error for HRES. On 16 March all members converged to the warm anomaly; the verification reached the 99th percentile of the model climatology. The large ensemble spread until 16 March indicated a large uncertainty in the forecast until this point. For most of the forecasts the HRES was among the best members but happened to have similar error magnitude as the worst members for 15 March 00UTC.

Figure 2 shows maps of z200 for the HRES forecast (black line), analysis (red line) and forecast error (shaded) for the forecast from 15 March 00UTC. This type of figure can be used for manual error tracking backwards in time. The 500 hPa level has also been used but is not shown here. The 6-day HRES forecast from 15 March 00UTC predicted a ridge over north-western Europe and a trough over eastern Europe, while the resulting analysis was much more zonal over western Europe and had the trough further to the east compared to the forecast. Overall the structure of the error in the HRES forecast had the signature of a westward phase shift of 15° longitude of the geopotential height pattern. Tracing the error backward in time is straightforward until the 3-day forecast (72h), which had a large error in three nodes streching from Mexico to north-eastern U.S/Canada. For shorter lead-times it is more complicated to track the

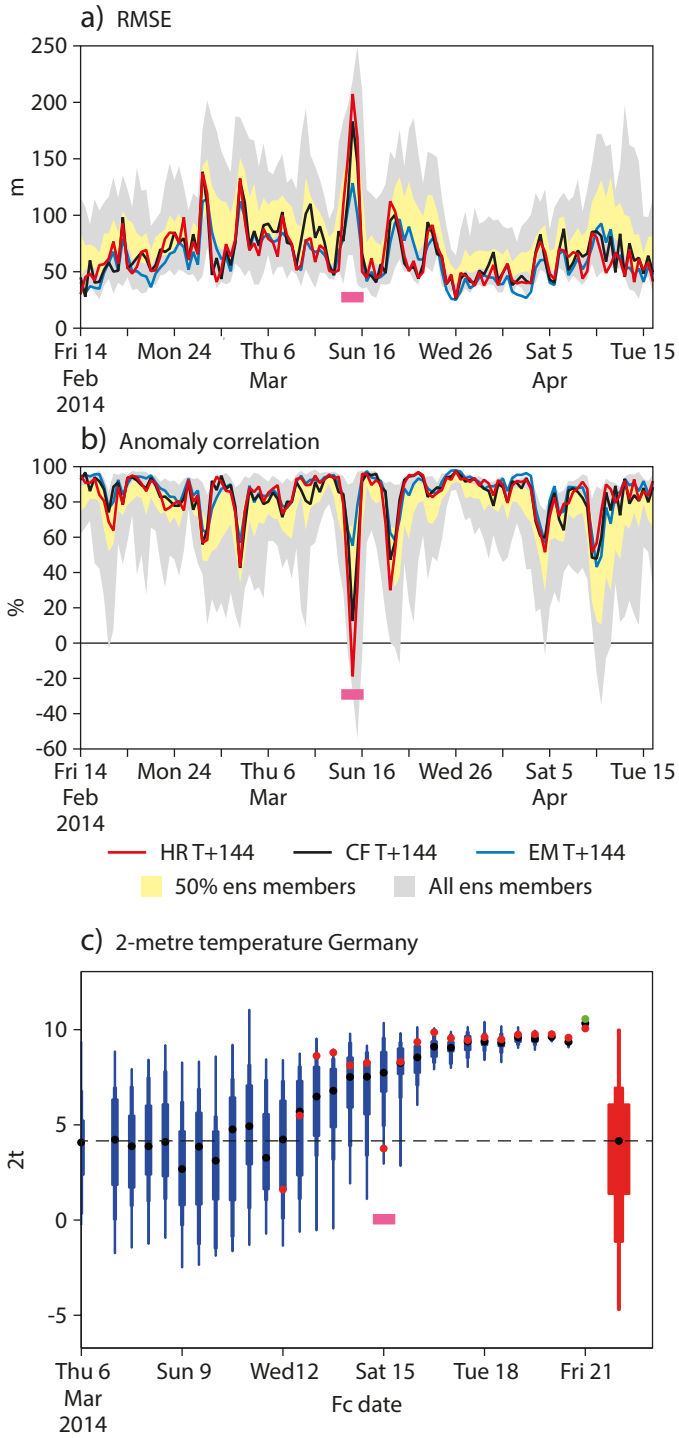

Figure 1: Time-series of root-mean-square-error (RMSE, top-left) and anomaly correlation coefficient (ACC, top-right) for 500 hPa geopotential height (z500) for 6-day forecasts for Europe. High-resolution forecast (HRES, red line), ensemble control (black line), ensemble mean (blue line) and the distributions of errors in individual ensemble members (grey - all members, yellow - 25 - 75 percentile). Daily mean 2-metre temperature on 21 March 2014 over Germany (bottom) for analysis (red), HRES (red) and ENS median (black dot) and probability boundaries for max-min, 10-90th and 25-75th percentile (blue box-and-whisker). The panel also include the distribution of the model climate (red box-and-whisker), with climate median as black line. The pink bar highlights the forecast investigated in the case.

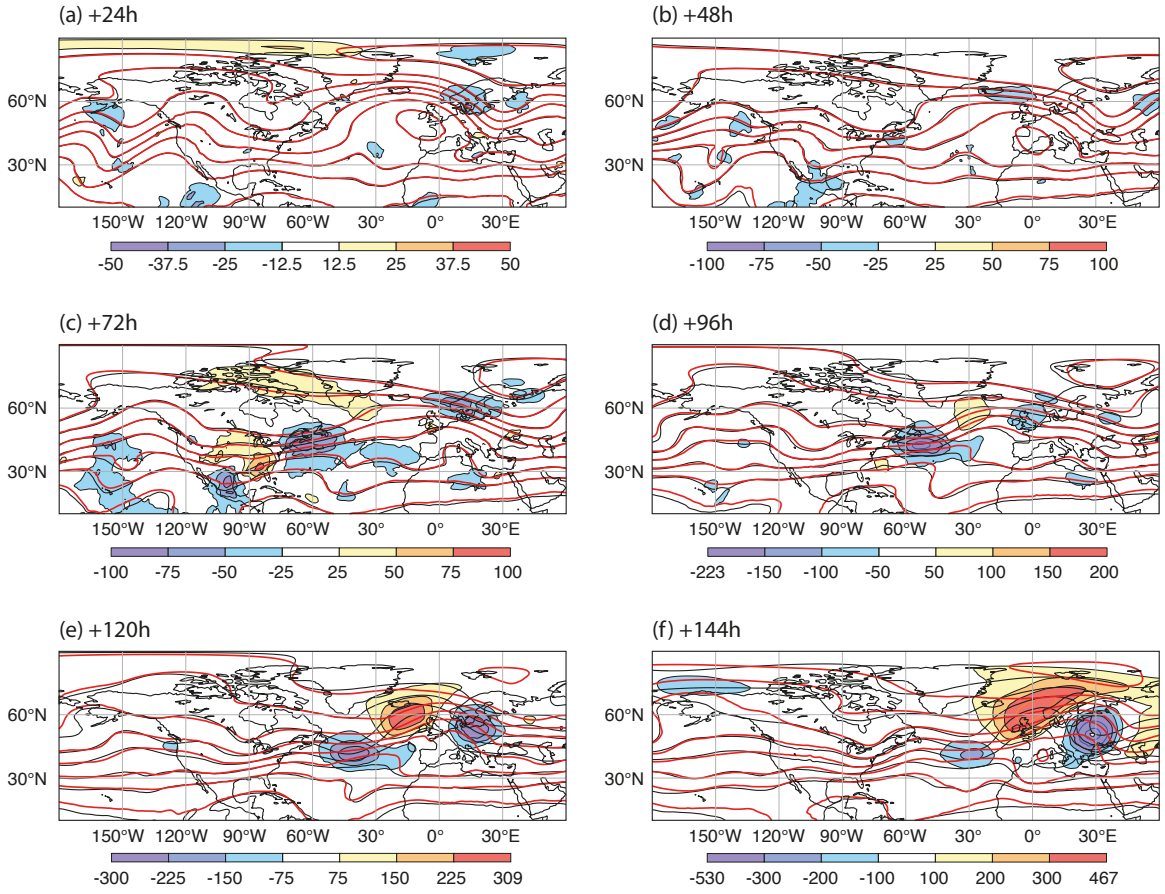

Figure 2: Forecasts from 00UTC 15 March 20014 of z200 for Case 1 and every 24 hour for the HRES (black line), verifying analysis (red line) with 200 metres contour interval and the forecast error shaded).

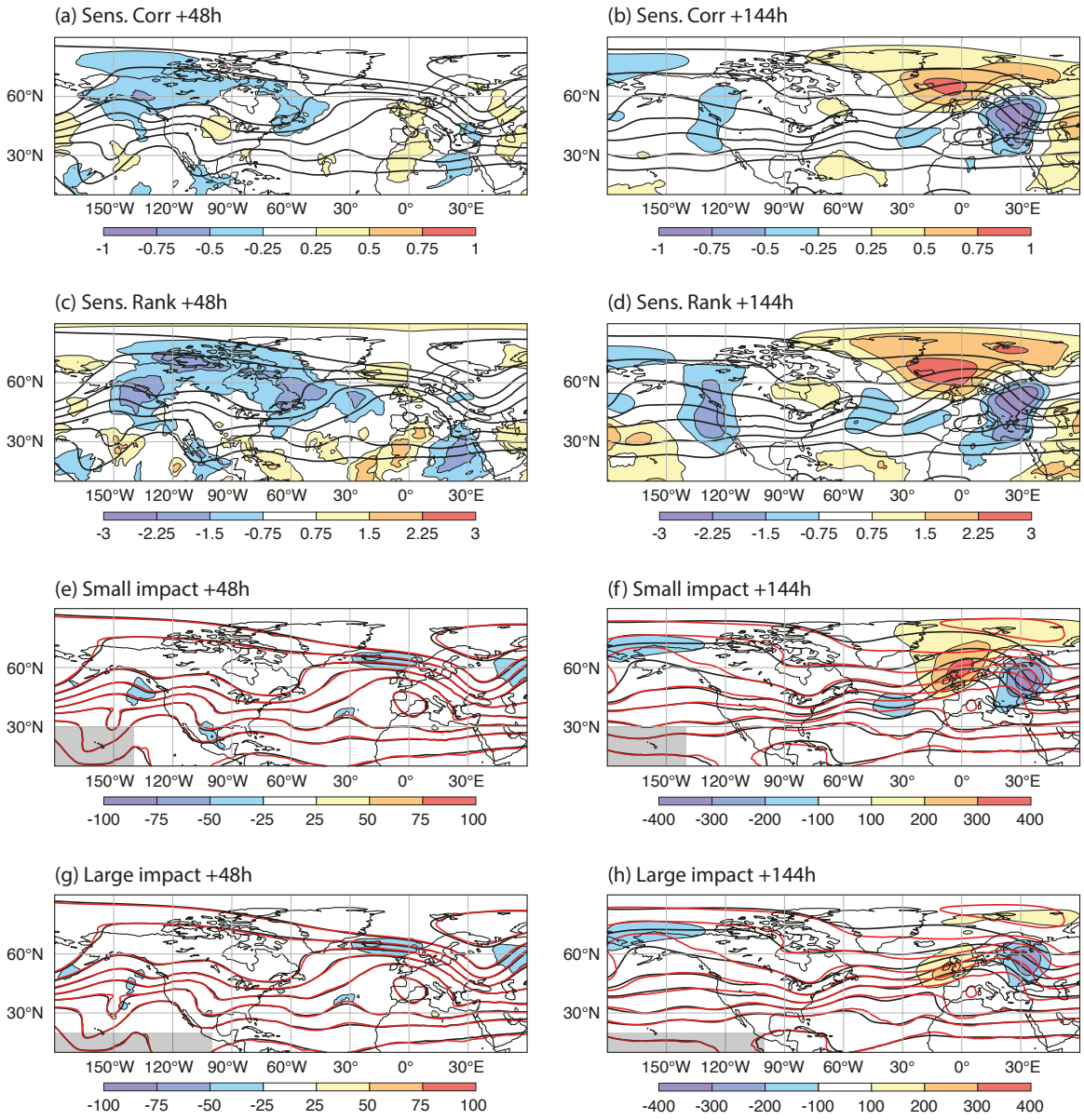

Figure 3: Forecast sensitivity (shading) for RMSE over Europe day 6 using the correlation method (a-b) and rank method (c-d). The contours in (a-c) represents ensemble mean of z200 on Day 2 and Day 6 respectively. Ensemble mean for z500 on day 6 (black), analysis (red) and ensemble mean error (shade) for relaxation experiment with small impact (e-f) and with large impact (g-h). The grey shaded area mark the relaxation box for the experiment. Diagnostics for Day 2 (left) and Day 6 (right). All diagnostics are for z200.

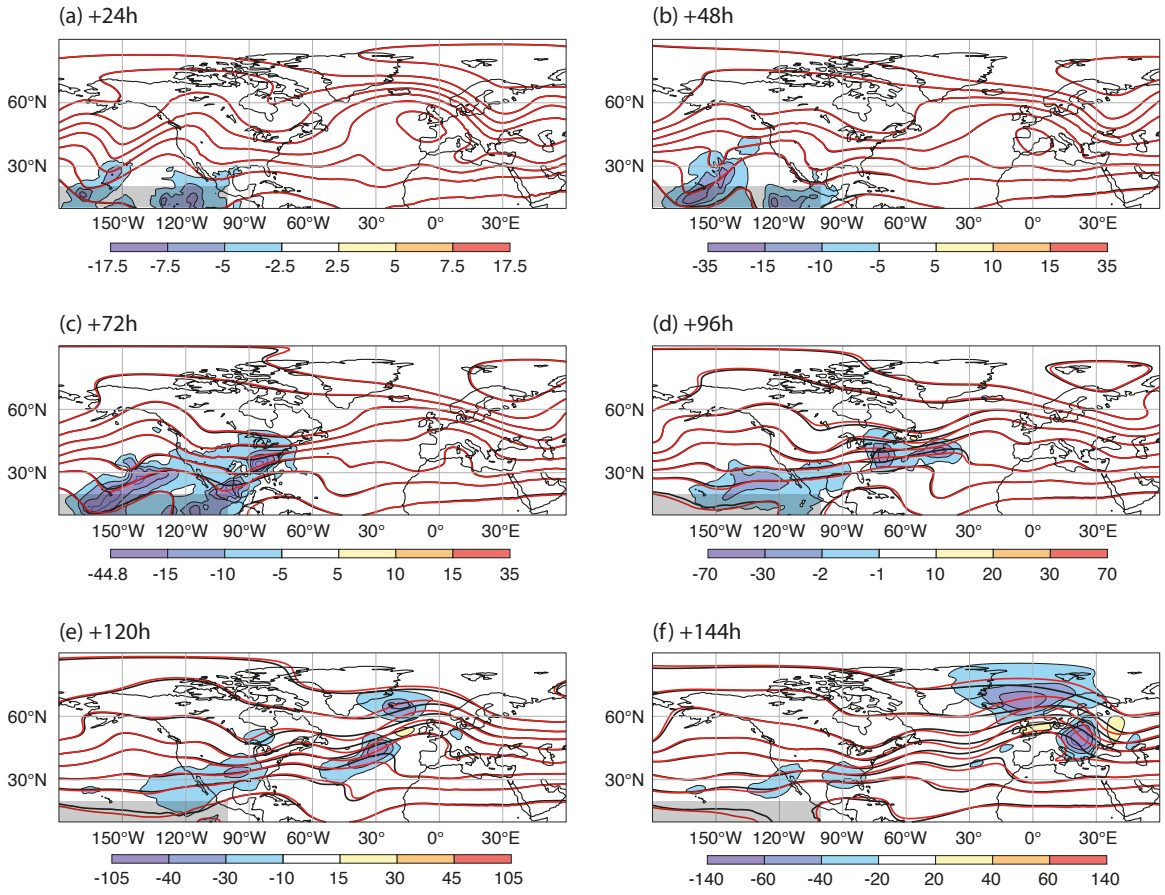

Figure 4: Ensemble mean for z200 from relaxation experiment with large impact (black) and with relaxation (red) and the difference in ensemble spread (shade) between the experiments. The black grey shaded area marks the relaxation box.

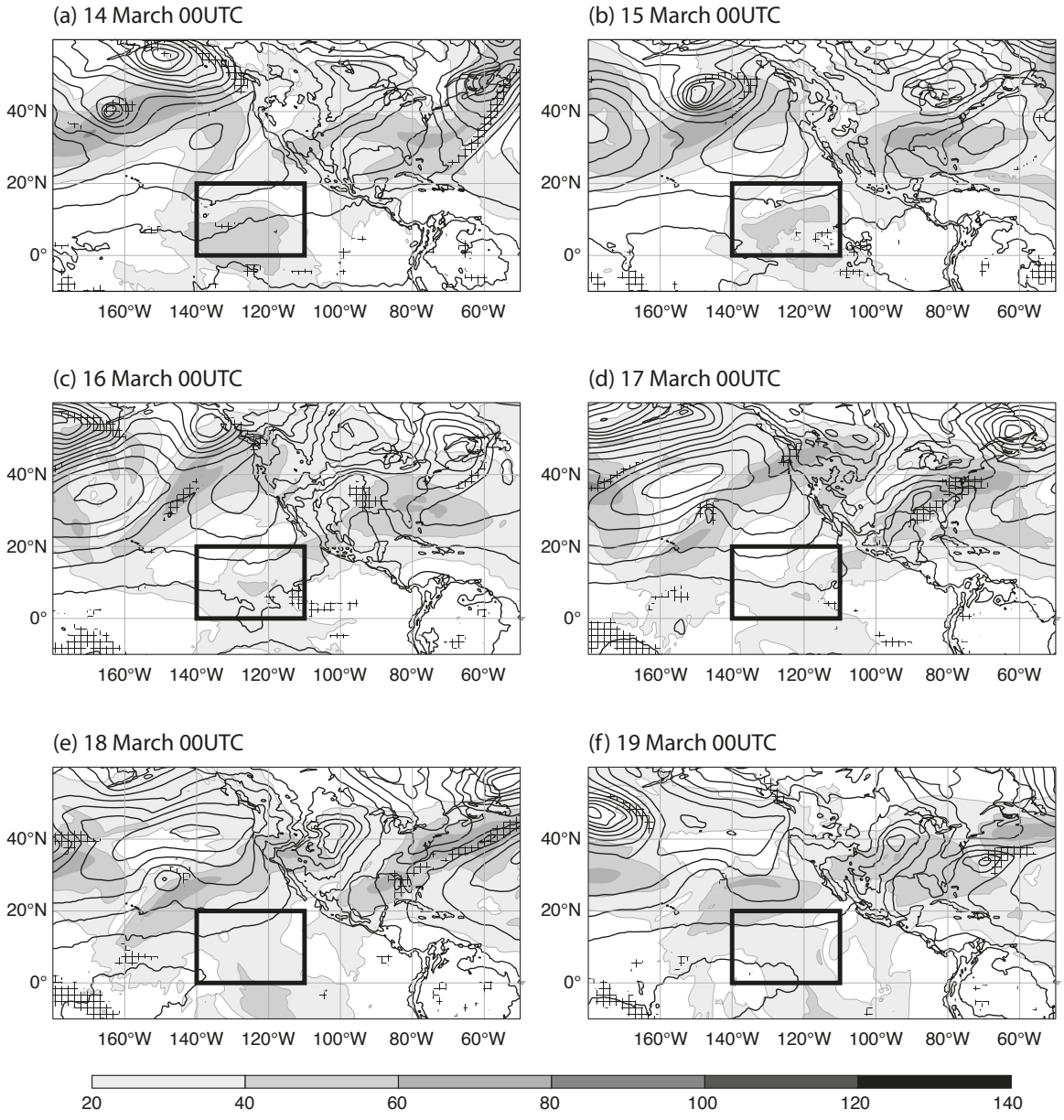

Figure 5: Analyses of MSLP (contour with 5hPa interval), wind speed at 200 hPa (shade) and areas with more than 5 mm precipitation in 6 hours (hatched) valid for Case 1. The suspected error source region (black box).

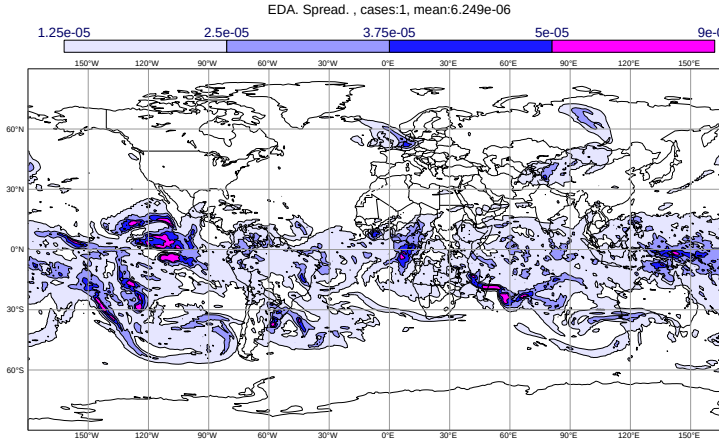

Figure 6: Standard deviation from Ensemble of data assimilations for vorticity on 200 hPa valid 15 March 00UTC.

error, but we find errors south-west of Mexico over the tropical Pacific. Regarding the 2-day forecast at z500 (not shown), errors were present with similar magnitude in several places, such as in a small-scale low over the North Atlantic, in a large-scale trough over eastern Canada and in a narrow trough over southern United States.

Figure 3 shows results from ensemble sensitivity(a-d) analysis and relaxation experiment(e-h). The ensemble sensitivity (using the z500 6-day error over Europe as response function) is shown for z200 at Day 2 (left) and Day 6 (right). The top row shows the sensitivity using the correlation method while the second row shows the sensitivity using the rank method as introduced in Section 2.1. Studying the results for Day 6, the correlation and rank method give similar results, which is expected as they are based on the same ensembles. The pattern of the sensitivity on Day 6 is similar to the error in HRES on Day 6, which is expected as the error for this lead time was used as the metric for the sensitivity. However at Day 2, it is difficult to find any outstanding structure in ensemble sensitivity method for this case.

Figure 3 also includes the 2-day (left) and 6-day (right) error for z200 ensemble mean from experiments with relaxation with small impact (e-f) and relaxation experiment with a significant reduction of the error over Europe (g-h). The grey shaded pattern shows the relaxation box. The relaxation box was chosen following the result of the error tracking where 1-day errors appeared south-west of Mexico. Comparing the two experiments, we find an impact from the relaxation area over the sub-tropical Pacific for the 6-day error over Europe for z200, and the reduction in error is even stronger for z500 (not shown). In particular the error in the ridge over north-western Europe is clearly reduced by applying the relaxation. Two days into the forecasts, the errors in z200 over Mexico are reduced in the relaxation experiment, which in the 3-day forecast strongly reduced the error in the two southerly error nodes seen in Figure 2(c) (not shown). The northerly node is still present in the relaxation experiment and might explain the remaining error at Day 6.

As the relaxation experiment is run as an ensemble, we also expect the relaxation to constrain the ensemble members inside the relaxation region. Figure 4 shows the difference in ensemble standard deviation (spread) between the experiment with and without relaxation. As expected we find decreased spread inside the relaxation box. The figure illustrates how this reduction is propagating downstream of the box with increased forecast lead time. After one day (a) the relaxation experiment has reduced the spread

over the central Pacific and also inside the eastern part of the relaxation box. The latter led to a reduction in spread over Mexico and the southern United States at Day 2 and 3, in line with the reduction in forecast error seen in Figure 3(d,f). Six days into the forecast we find a reduction of the ensemble spread by the relaxation over Europe. This confirms that a reduction in the ensemble spread (by contraining the ensemble members) in the box impacted the spread over Europe on a 6-day time scale.

To further understand the flow situation in the suspected source region for the error, Figure 5 shows a sequence of analyses every 24 hours, starting at 14 March 00UTC. Each panel includes MSLP (black contour), wind speed at 200 hPa as a proxy of the jet stream (shade) and areas with more than 5 mm precipitation in 6 hours (hatched). The region of the suspected initial error is marked with a black box. As seen in Figure 5, the jetstream reached as far south as the equator in the eastern Pacific around 15 March. The flow pattern resembles the structure of an equatorial Rossby wave, determined by a modal decomposition of the flow following Zagar *et al.* (2015) (not shown). On 17 March the jet streak (core of the jetstream) advanced towards Mexico and Texas. On 18 March severe convection appeared over southern-eastern U.S and on 19 March the jet streak influenced the cyclogenesis over north-eastern U.S.

At the initialisation time of the last poor forecast during this bust (15 March 00UTC), large spread was present in the ensemble of data assimilation (EDA) for upper-tropospheric winds in the area of suspected initial error (Figure 6), which indicates large initial uncertainties. On average, this is a problematic region for the forecast system, with among the largest wind-errors at 200 hPa in short-range forecasts around the globe (not shown). The eastern tropical Pacific is an area with very few conventional wind observations and the main source of wind observations is from atmospheric motion vectors, which builds on assumptions such as the height of the tracked clouds. For this case we suspect that an initial error in the upper-level winds affected the jet-streak towards the southern U.S. The error later amplified during interaction with convection over the southern U.S and the cyclogenesis later on.

3.2 Case 2 - 20 October 2015 12UTC

The second case presented here is from the end of October 2015. The bust for 6-day forecasts was present in relatively few consecutive forecasts. Very large errors, especially in terms of ACC (Figure 7(b)), for the HRES forecasts appeared on 20 October (both 00UTC and 12UTC). However, the uncertainty in the 6-day forecasts remained until 23 October 00UTC with more than 25% of the ensemble members having ACC below 20%. For the worst forecast (20 October 12UTC) the ACC for ensemble mean went down to -20% and no member had better ACC than 30%. It is very unusual that all ensemble members show such large errors.

Figure 7(c) shows the evolution of the ensemble forecast and HRES for the 2-metre temperature over Germany for 28 October 2015, 7.5 days into the worst forecast for this case (instead of 6 days in the similar figures for Case 1 and 3). The forecasts from around 20 October 2015 predicted colder temperatures than normal, and especially for the forecast from 20 October 12UTC almost 90% of the ensemble members were colder than the model climate, while the outcome was a slight warm anomaly. On the 21 October, the ensemble and HRES changed towards normal conditions and on 22 October 00UTC the ensemble spread was much reduced.

The error in the 6-day forecast was associated with the failure to predict a ridge over north-western Europe, as seen in Figure 8(f). Tracking the error in HRES backward is straightforward until Day 3 where we find large errors in z200 associated with a ridge of central Canada and a trough over further east. One day earlier we find small errors related to a trough over the Great Lakes and also in a ridge over northern Canada.

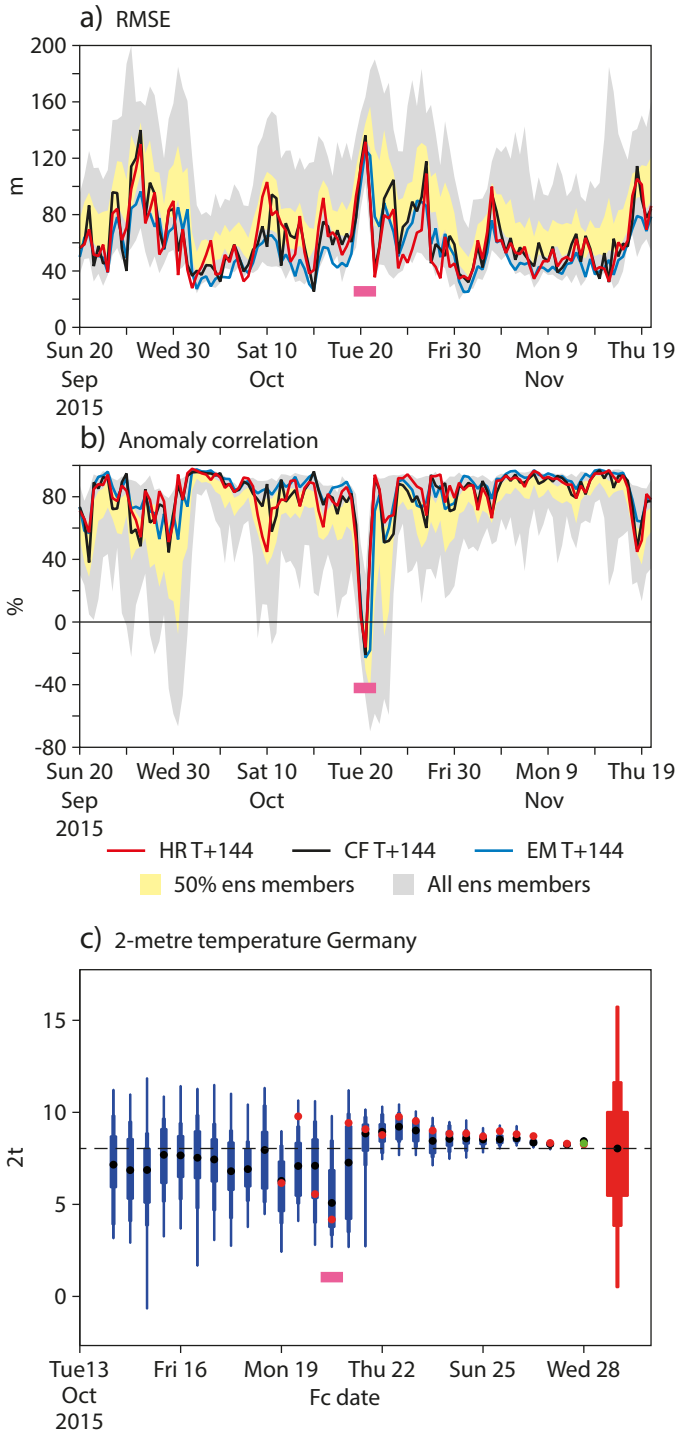

Figure 7: Same as Figure 1 but for Case 2.

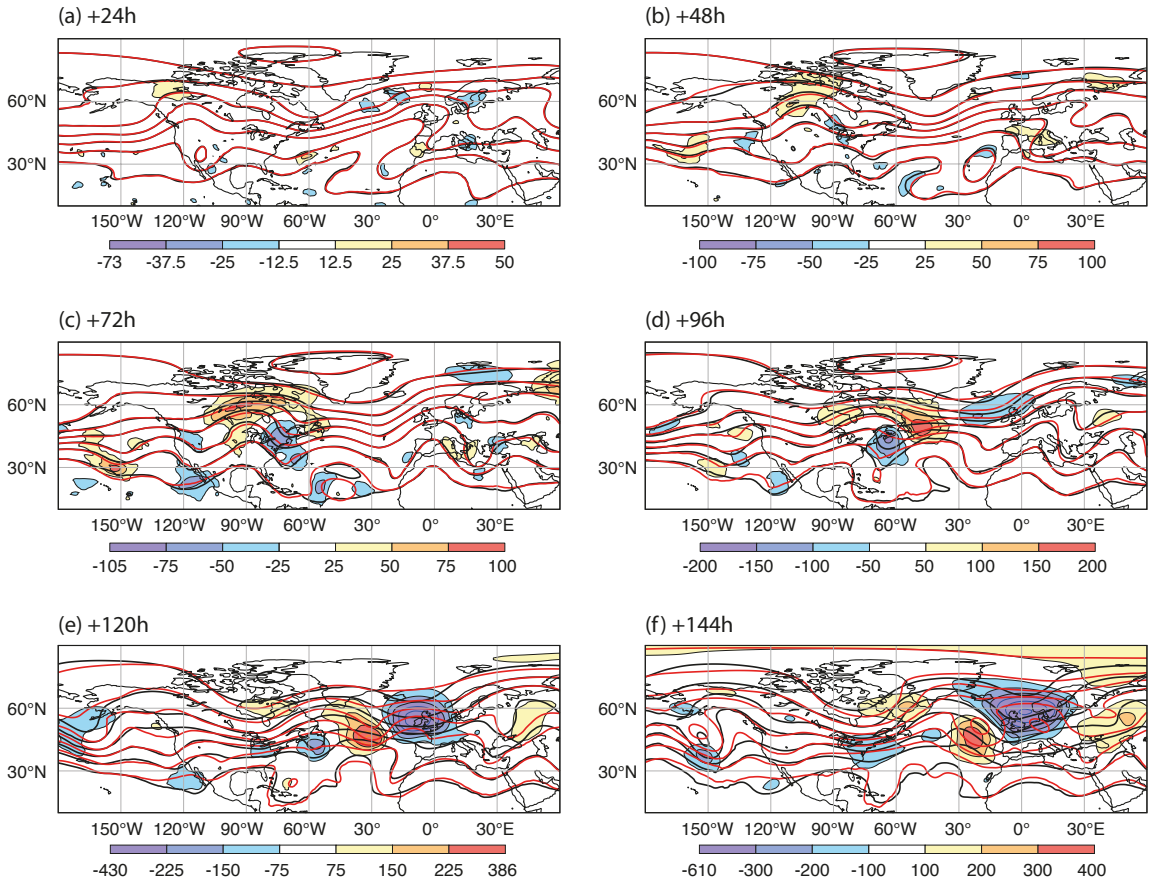

Figure 8: Same as Figure 2 but for HRES forecast from 20 October 12UTC.

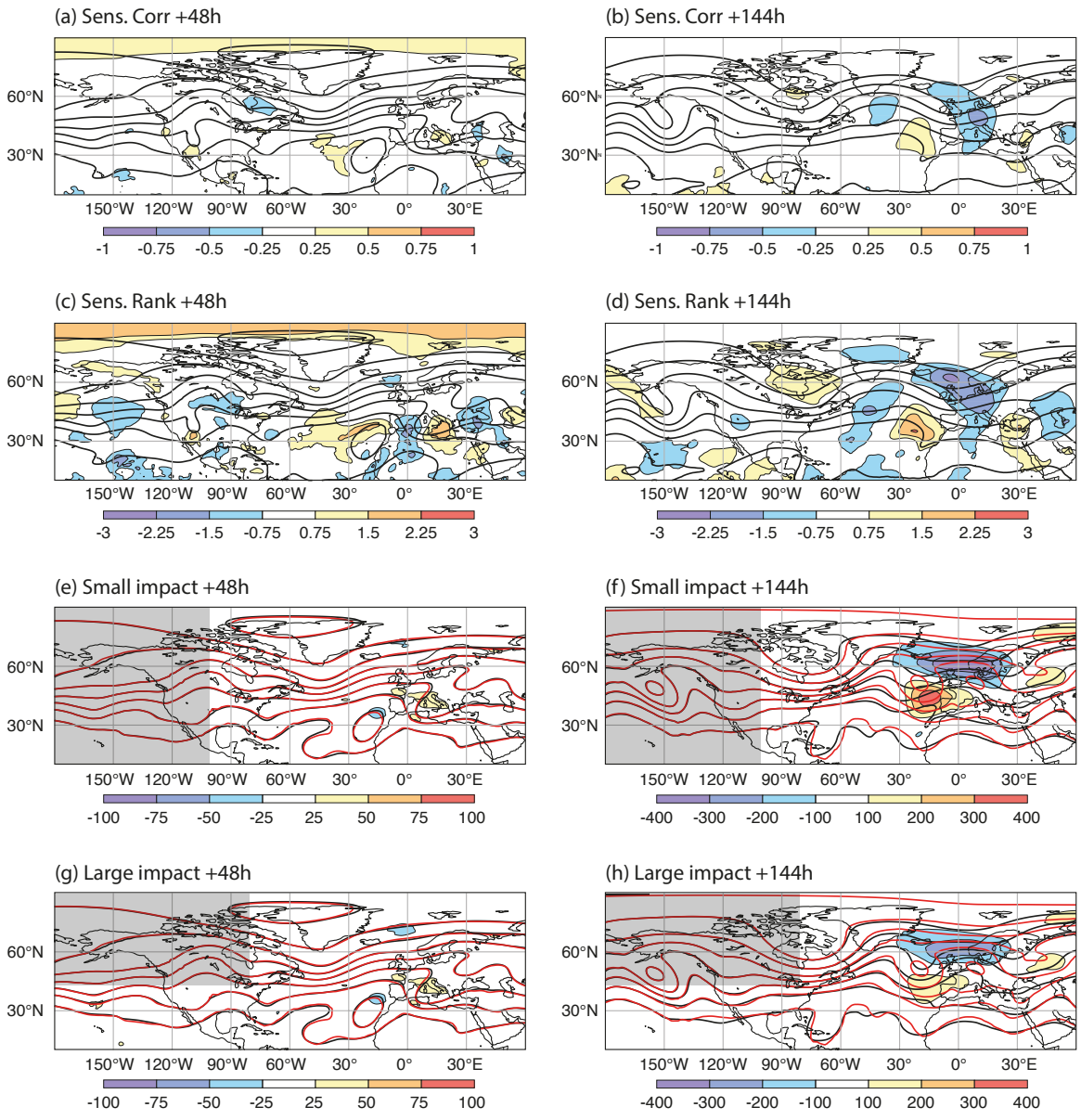

Figure 9: Same as Figure 3 but for Case 2.

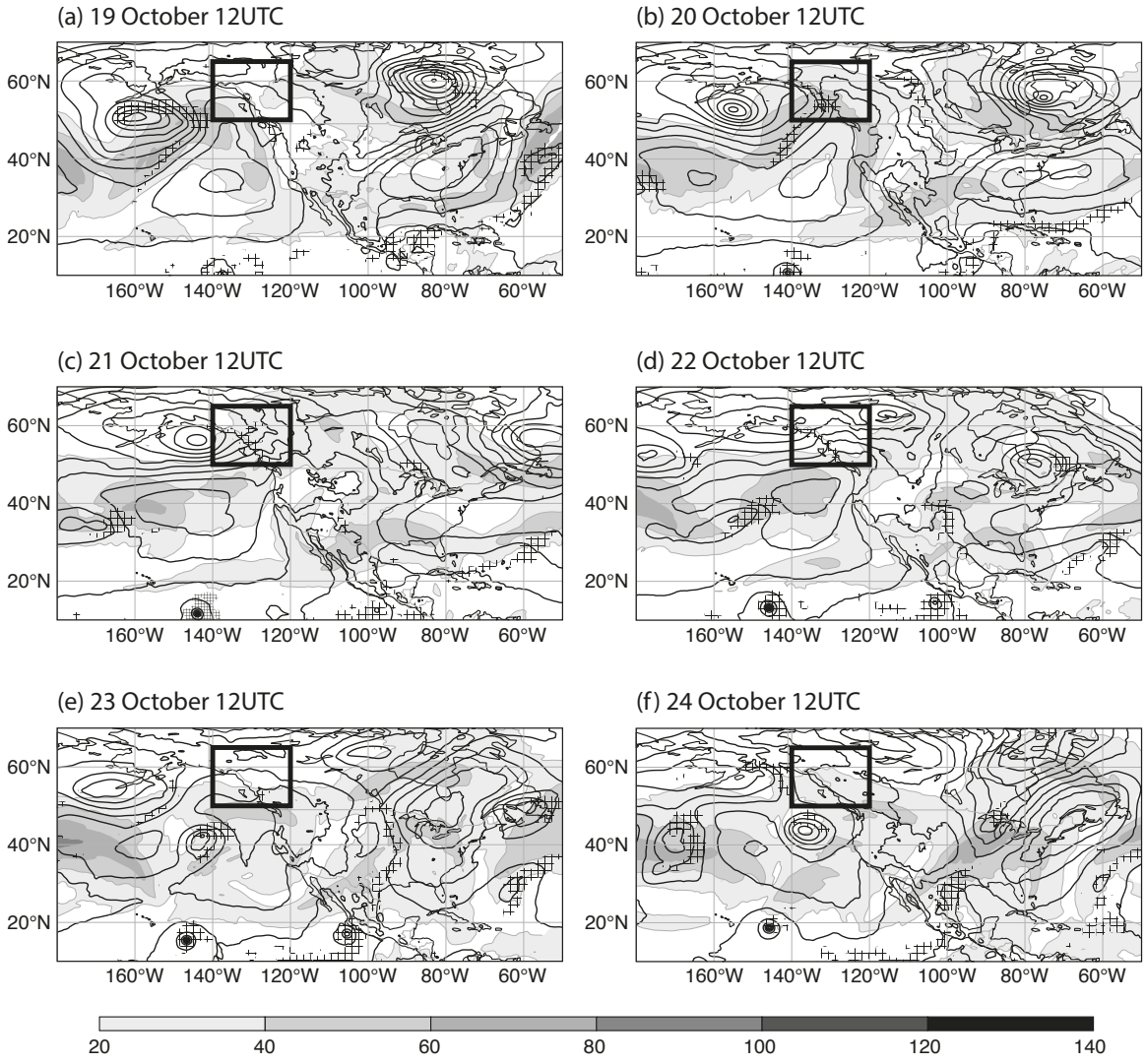

Figure 10: Same as Figure 5 but for Case 2.

Studying the results from the ensemble sensitivity analysis (Figure 9(a-d)), we also find some weak sensitivites at Day 2 in the same region using the rank method, but there are many different structures in the plot and the structure around the Great Lakes is not significantly different. However, it is no surprise that the ensemble sensitivity method gives very little information for this case as all ensemble members contained large errors, and hence did not capture the true evolution of the atmosphere.

For the relaxation, we compare two experiments where the broadening in relaxation regions is the inclusion of central Canada (Figure 9), which reduces the error over Europe on Day 6. The relaxation probably reduced the errors in the forecasts before the cyclogenesis north of the Great Lakes (Figure 10(d)), but the cyclogenesis still created further uncertainties and the error in the relaxation experiment is still relatively high. An additional experiment including the area of the cyclone development further reduced the error (not shown). The rapidly amplifying error associated with the developing cyclone could be related to model error or a small existing error that quickly amplified by the intrinsic uncertainty. However, as the first good forecast was produced from an analysis (21 October 00UTC) before the cyclongenesis was observed, it suggests that the error was present before the cyclone developed and not only caused by model error related to the formation of the cyclone. Anyway, large uncertainties remained in the ensemble forecast until the cyclone was present in the analysis (23 October 00UTC).

3.3 Case 3 - 7 March 2016 00 UTC

The errors in the forecasts from the first week in March 2016 were associated with the onset of strong blocking over Scandinavia. The first forecast with large 6-day forecast errors for Europe was 4 March 12 UTC and consecutive forecasts were affected until 7 March 00UTC (Figure 11,a-b). The forecasts from this period predicted warmer than normal temperatures over Germany, while the outcome was normal temperatures. The ensemble converged to the true solution between the forecast from 7 March 00UTC and 8 March 00UTC.

For this case the error propagation is more clear in z500 than in z200, and Figures 12 and 13 show this level instead. The forecast from 7 March 00UTC experienced the largest 4-day RMSE over Europe at since 2001 for the ECMWF HRES forecast. The nature of the 6-day error was an underestimation of the north-eastward extent of a ridge that built up over Scandinavia, which in turn led to an overestimation of the temperatures over Germany as discussed above. The forecast captured the ridge to some extent but clearly underestimated the amplitude. In the 2-day forecast we find large errors in the ridge between Greenland and Iceland (underestimated in the forecast), together with errors in the trough further upstream. Already in the 1-day forecast, clear error structures are present over the western Atlantic with one negative node east of New Foundland and one postive node further south.

The Day 2 ensemble sensitivities (Figure 13) have a strong structure over the western Atlantic, which is very similar to the structure of the day 2 error in HRES. This strengthens the hypothesis of the importance of this structure. The senstivity to this region is also clear in the relaxation experiments. By nudging the forecast from the dateline to 75°W, very little impact is found for the Day 6 error over Europe, while the experiment with the relaxation region extended to 60°W had a large positive impact on the error.

Regarding the synoptic situation (Figure 14), a surface low, squeezed between two anticyclones, developed in the trough on 7 March (Figure 14). The low developed over the Gulf Stream with cold air heading south on its westward side (at about 60°W), leading to strong upward surface heat-flux. The last poor forecast (7 March 00z) had the cyclone 24-hours into the forecast 4° too the south compared to the analysis. Forecasts from 5 and 6 March showed a similar error and it was not until the development of the cyclone was finally observed that the forecast improved.

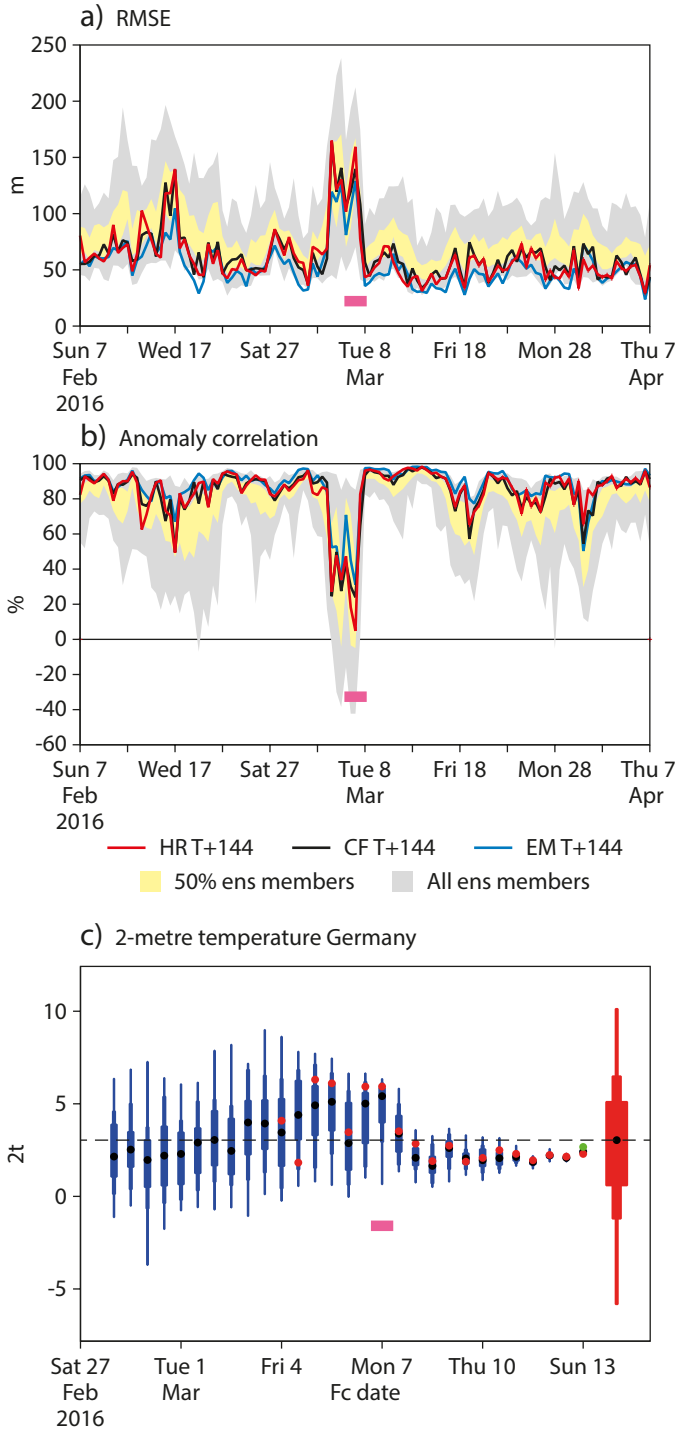

Figure 11: Same as Figure 1 but for Case 3.

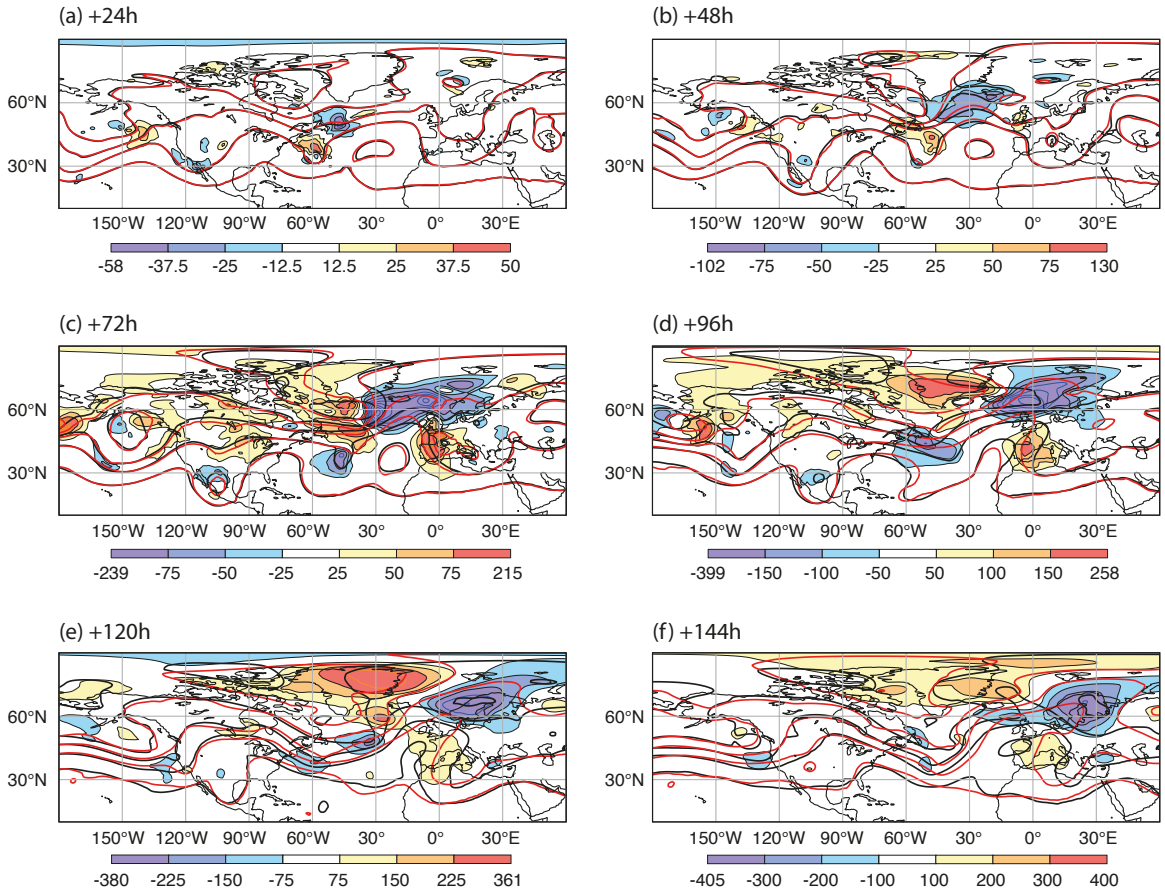

Figure 12: Same as Figure 2 but for HRES forecast from 7 March 00UTC and for z500 instead of z200.

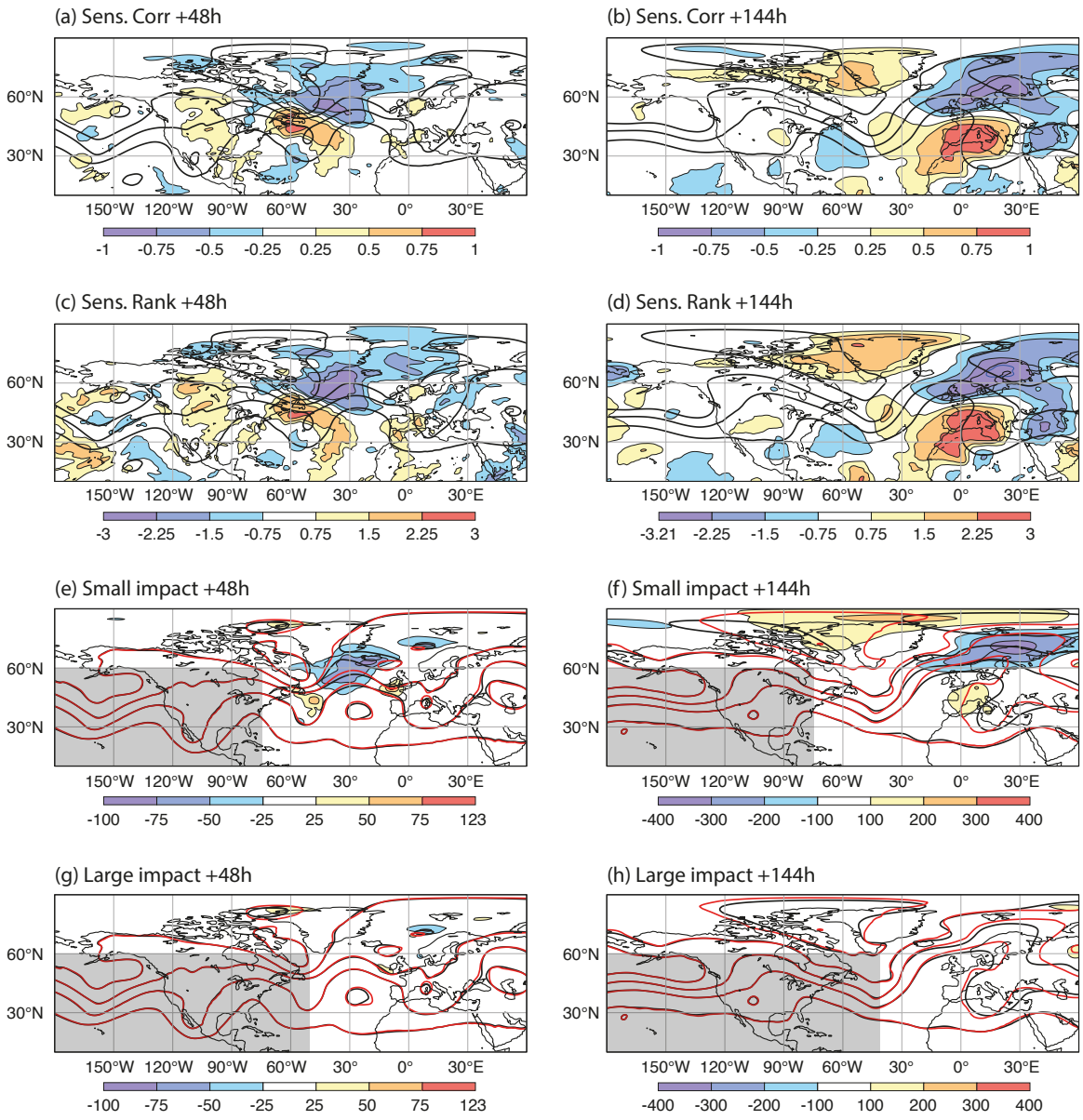

Figure 13: Same as Figure 3 but for Case 3 and forecasts of z500.

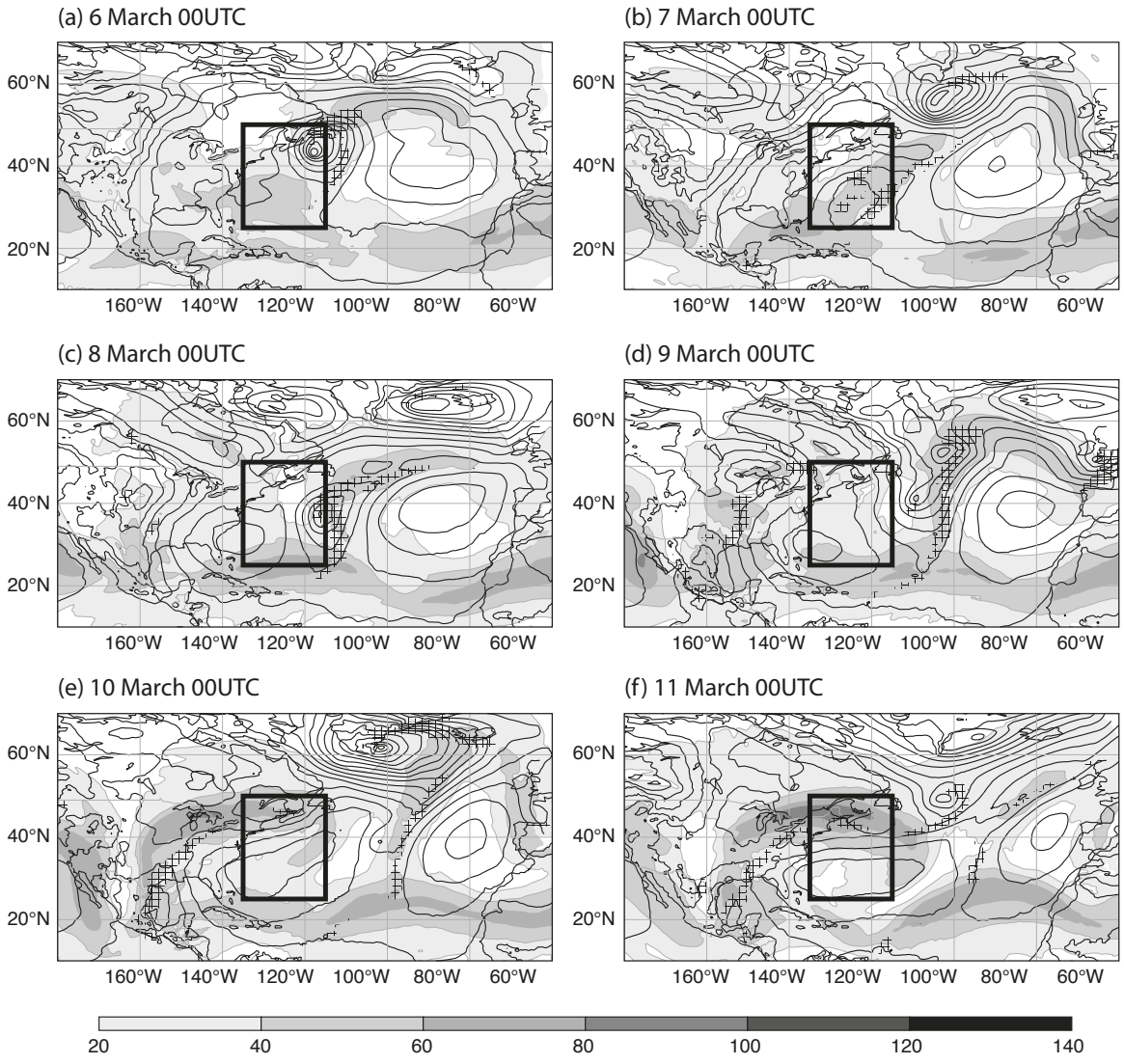

Figure 14: Same as Figure 5 but for Case 3.

4 Discussion

In this report a combination of three error tracking methods has been used to better understand three cases of low forecast skill (busts) over Europe. These cases illustrate different structures, origins of forecast errors and propagation speeds.

The most basic method to find the origin of the forecast error is to manually track error structures backward in time, starting from the final error (at Day 6 for example). In the ideal case one finds a large analysis increment (or error in 12-hour forecast) in the beginning of the error trajectory, which would point to the change needed by the assimilation to correct the forecast. However, the downside with the method is the difficulties that arise if the error splits into several structures. One also needs to be aware that the error often propagates with the group speed of the Rossby wave, and therefore can 'jump' between ridges and troughs of the wave package. In clear cases like Case 3, only using error tracking would have given a strong indication of the source region of the error, while in a situation like Case 1 and 2, it would have been more difficult to interpret the result from the method.

Ensemble sensitivity techniques build on the idea of comparing good and bad ensemble members. This is a simple technique to apply if the ensemble already exists with a sufficient number of ensemble members. The first requirement to be successful with the technique is to have some ensemble members that captured the situation with small errors. In Case 2, all ensemble members failed to predict the situation and hence the ensemble sensitivity will probably not give a clear answer. The results here indicate that the rank method give stronger signals but the results are more noisy.

The relaxation technique uses information of the truth (here with the analysis as pseudo-truth), to constrain the forecast in a specific region. If the error propagated through (or originated from) the region, the forecast would be improved if the analysis used was free from the key errors. The technique is a way to confirm the downstream impact from suspected regions found using the methods above. However, to apply this technique, access to the model is needed and it needs to have a nudging code available. This puts strong constraints on the possibility of using this method. The method is also time-consuming as one needs to guess the size of the region and try to find a region as far upstream as possible. Here the error tracking and ensemble sensitivity are useful to give a first guess. For the studied cases here, the manual tracking gave a clearer guidance to select the relaxation area than the ensemble sensitity methods. If successful, the relaxation method will give a strong indication of the source of the key error.

Another very useful method that we have not included in this report is to compare forecasts from different centres. For this, the TIGGE archive of global ensemble forecast (Bougeault *et al.*, 2010) is of much use. These data were, for example, used for this purpose in Rodwell *et al.* (2013) and recently in Yu and Meng (2016). If one or more of the forecasting centres had a significantly better prediction, one could look further into the forecast and initial condition differences. In Rodwell *et al.* (2013) this was taken one step further and initial conditions were swapped between the ECMWF and UKMO models to see the relative influence of initial and model errors for the case. However, this requires access to different models and infrastructure to start the forecast with initial conditions from a different model.

Apart from swapping initial conditions between models, it is very difficult to disentangle initial error and model errors. A further complication is that the initial errors often are due to errors in the first guess forecasts used in the data assimilation, and the initial condition error could hence be due to model errors. It is also not certain that the final error is a product of one source, but a superposition of different sources of errors. For Case 1, some error remains also in the relaxation experiment, which could be related to additional error structures that originated from further north.

Once the synoptic feature related to the error has been identified, additional diagnostics can be undertaken

to further understand the situation and the origin of the error. If there is a strong indication of the forecast error originating from initial error, it is of interest to see what type of observations that was able to correct the error. For this task observation statistics are essential and Forecast Sensitivity Observation Impact (Cardinali, 2009) can give guidance. If one suspects the model errors to be behind the forecast error, model tendency diagnostics (Klinker and Sardeshmukh, 1992; Rodwell and Palmer, 2007) can be useful to understand the active processes at the time of the appearance of the error.

In Rodwell *et al.* (2013) and recently in Lillo and Parsons (2016), European forecast bust cases from 1990 to 2010 extracted from ERA-Interim forecasts, were investigated in a climatological perspective. The bust cases connected to flow-regime transitions over the Euro-Atlantic region, especially related to blockings. For the cases presented here, the errors are also related to the Scandinavian blocking pattern, either underpredicted in the forecast (Case 2 and 3) or falsely predicted (Case 1). Verification also shows that the Scandinavian blocking is the most difficult to predict of the four Euro-Atlantic regimes (Ferranti *et al.*, 2015). For example Phahl *et al.* (2015) showed the importance of latent heat release in ascending air streams (often referred to as Warm Conveyor Belts) for blocking cases, and ongoing work will try to relate the errors in the cases presented here to warm-conveyor belt occurence.

Another source region for errors worth more exploration is the tropics. In Case 1 the key error was found to originate from the eastern Tropical Pacific during a very high amplitude episode of the sub-tropical jetstream reaching the equator. A way to explore the predictability in the tropics is to evaluate atmospheric modes as in Zagar *et al.* (2015). It is also believed that an active phase of the Madden-Julian oscillation in the central Pacific can trigger the onset of negative a NAO pattern (Cassou, 2008).

It is of importance to understand the dynamics of the intrinsic amplification of errors. In Case 1, a small existing error in the jet stream amplified over an area with intense convection and in Case 2 it amplified in connection to a developing cyclone. The effect of moist processes for intrinsic error growth has recently been discussed in Sun and Zhang (2016). As we are interested in longer and longer forecast ranges, the error will propagate through several episodes of high intrinsic error growth. It is also worth highlighting that none of the cases here is related to extra-tropical transition of tropical cyclones. These are known as a large source of uncertainty for medium-range errors over Europe (Jones *et al.*, 2003; Grams *et al.*, 2013). In Lillo and Parsons (2016) it was suggested that this is the main source for forecast busts over Europe in the autumn.

A natural step to extend this study is to use the methods to calculate a climatology of sources for forecast errors, in a similar way as in Rodwell *et al.* (2013) and in Lillo and Parsons (2016). For this purpose the ensemble sensitivity method would be the most suitable as it does not require additional simulations. However, the question is how far back in lead-time one can reach significant results and how to filter spurious correlations.

Finally, all the cases presented here showed fast error growth due to instabilities in the atmospheric flow. In some cases large errors are inevitable due to the intrinsic growth of error in the atmosphere. Therefore one has to appreciate that the atmopshere sometimes is unpredictable and even a perfect model with near-perfect initial conditions will produce an unskilful forecast. However, the unpredictable situations should be foreseen by a large spread in the ensemble in order to have a reliable forecast system.

Acknowledgements

I would like to acknowledge Christian Grams at ETH Zürich and David Richardson, David Lavers and Mark Rodwell at ECMWF for valuable comments. I would also like to thank Anabel Bowen for help

with the preparation of the figures.

References

Ancell B, Hakim GJ. 2007. Comparing adjoint- and ensemble-sensitivity analysis with applications to observation targeting. *Monthly Weather Review* **135**(12): 4117–4134, doi:10.1175/2007MWR1904.1.

Bauer P, Thope A, Brunet G. 2015. The quiet revolution of numerical weather prediction. *Nature* **525**: 47–55.

Bougeault P, Toth Z, Bishop C, Brown B, Burridge D, Chen D, Ebert E, Fuentes M, Hamill T, Mylne K, Nicolau J, Paccagnella T, Park YY, Parsons D, Raoult B, Schuster D, Silva Dias P, Swinbank R, Takeuchi Y, Tennant W, Wilson L, Worley S. 2010. THORPEX Interactive Grand Global Ensemble (TIGGE). *Bull. Amer. Meteor. Soc.* **91**: 1059–1072.

Cardinali C. 2009. Monitoring the observation impact on the short-range forecast. *Q. J. R. Met. 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.

Ferranti L, Corti S, Janousek M. 2015. Flow-dependent verification of the ecmwf ensemble over the euro-atlantic sector. *Quarterly Journal of the Royal Meteorological Society* **141**(688): 916–924, doi:10.1002/qj.2411, URL http://dx.doi.org/10.1002/qj.2411.

Grams CM, Jones SC, Davis CA. 2013. The impact of typhoon jangmi (2008) on the midlatitude flow. part ii: Downstream evolution. *Quarterly Journal of the Royal Meteorological Society* **139**(677): 2165–2180, doi:10.1002/qj.2119.

Jones SC, Harr PA, Abraham J, Bosart LF, Bowyer PJ, Evans JL, Hanley DE, Hanstrum BN, Hart RE, Lalaurette F, Sinclair MR, Smith RK, Thorncroft C. 2003. The Extratropical Transition of Tropical Cyclones: Forecast Challenges, Current Understanding, and Future Directions. *Wea. Forecasting* **18**: 1052–1092.

Jung T. 2011. Diagnosing remote origins of forecast error: relaxation versus 4d-var data-assimilation experiments. *Quarterly Journal of the Royal Meteorological Society* **137**(656): 598–606, doi:10.1002/qj.781, URL http://dx.doi.org/10.1002/qj.781.

Jung T, Kasper MA, Semmler T, Serrar S. 2014. Arctic influence on subseasonal midlatitude prediction. *Geophysical Research Letters* **41**(10): 3676–3680, doi:10.1002/2014GL059961, URL http://dx.doi.org/10.1002/2014GL059961.

Jung T, Miller M, Palmer T. 2010a. Diagnosing the origin of extended-range forecast errors. *Mon. Wea. Rev.* **138**: 2434–2446.

Jung T, Miller M, Palmer T. 2010b. Understanding the anomalously cold European winter of 2005/2006 using relaxation experiments. *Q. J. R. Met. Soc.* **116**: 3157–3174.

Kelly G, Thepaut JN, Buizza R, Cardinali C. 2007. The value of observations. i: Data denial experiments for the atlantic and the pacific. *Quarterly Journal of the Royal Meteorological Society* **133**(628): 1803–1815, doi:10.1002/qj.150, URL http://dx.doi.org/10.1002/qj.150.

Klinker E, Sardeshmukh PD. 1992. The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atm. Sci.* **49**: 608–627.

Lamberson WS, Torn RD, Bosart LF, Magnusson L. 2016. Diagnosis of the Source and Evolution of Medium-Range Forecast Errors for Extratropical Cyclone Joachim. *Weather and Forecasting* **31**(4): 1197–1214, doi:10.1175/WAF-D-16-0026.1.

Lillo SP, Parsons DB. 2016. Investigating the dynamics of error growth in ecmwf medium range forecast busts. *Quarterly Journal of the Royal Meteorological Society* : n/a–n/adoi:10.1002/qj.2938, URL http://dx.doi.org/10.1002/qj.2938. QJ-15-0196.R2.

Magnusson L, Källén E. 2013. Factors influencing skill improvements in the ECMWF forecasting system. *Mon. Wea. Rev.* **141**: 3142–3153.

Phahl S, Schwierz C, Croci-Maspoli M, Grams CM, Wernli H. 2015. Importance of latent heat release in ascending air streams for atmospheric blocking. *Nature Geoscience* **8**: 610–614.

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

Rodwell MJ, Palmer TN. 2007. Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society* **133**(622): 129–146, doi:10.1002/qj.23, URL http://dx.doi.org/10.1002/qj.23.

Sun YQ, Zhang F. 2016. Intrinsic versus Practical Limits of Atmospheric Predictability and the Significance of the Butterfly Effect. *Journal of the Atmospheric Sciences* **73**(3): 1419–1438, doi:10.1175/JAS-D-15-0142.1.

Torn RD, Hakim GJ. 2008. Ensemble-based sensitivity analysis. *Monthly Weather Review* **136**(2): 663–677, doi:10.1175/2007MWR2132.1.

Torn RD, Whitaker JS, Pegion P, Hamill TM, Hakim GJ. 2015. Diagnosis of the source of gfs medium-range track errors in hurricane sandy (2012). *Monthly Weather Review* **143**(1): 132–152, doi:10.1175/MWR-D-14-00086.1.

Yu H, Meng Z. 2016. Key synoptic-scale features influencing the high-impact heavy rainfall in beijing, china, on 21 july 2012. *Tellus A* **68**(0), doi:10.3402/tellusa.v68.31045.

Zagar N, Buizza R, Tribbia J. 2015. A three-dimensional multivariate modal analysis of atmospheric predictability with application to the ECMWF ensemble. *J. Atm. Sci.* **72**: 4423–4444.

Zheng M, Chang EKM, Colle BA. 2013. Ensemble sensitivity tools for assessing extratropical cyclone intensity and track predictability. *Weather and Forecasting* **28**(5): 1133–1156, doi:10.1175/WAF-D-12-00132.1.