University of Reading

Evaluation of the consistency of ECMWF ensemble forecasts

Article

Published Version

Creative Commons: Attribution 4.0 (CC-BY)

Open Access

Richardson, D. S., Cloke, H. L. ORCID: https://orcid.org/0000-0002-1472-868X and Pappenberger, F. (2020) Evaluation of the consistency of ECMWF ensemble forecasts. Geophysical Research Letters, 47 (11). e2020GL087934. ISSN 0094-8276 doi: 10.1029/2020GL087934 Available at https://centaur.reading.ac.uk/90485/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1029/2020GL087934

Publisher: American Geophysical Union

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

University of Reading

Central Archive at the University of Reading

Reading's research outputs online

Geophysical Research Letters

RESEARCH LETTER
10.1029/2020GL087934

Evaluation of the Consistency of ECMWF Ensemble Forecasts

David S. Richardson[1,2], Hannah L. Cloke[2,3,4], and Florian Pappenberger[1]

[1]European Centre for Medium-Range Weather Forecasts, Reading, UK, [2]Department of Geography and Environmental Science, University of Reading, Reading, UK, [3]Department of Meteorology, University of Reading, Reading, UK, [4]Department of Earth Sciences, Uppsala University, Uppsala, Sweden

Key Points:
- A new divergence index is introduced to measure inconsistency (jumpiness) in a sequence of ensemble forecasts
- The ECMWF ensemble has occasional large inconsistency between successive runs, with the largest jumps tending to occur at 7–9 days lead
- To understand the causes of jumpiness it is important to consider the time evolution of each ensemble (e.g., using phase-space trajectories)

Supporting Information:
- Supporting Information S1

Correspondence to:
D. S. Richardson,
david.richardson@ecmwf.int

Citation:
Richardson, D. S., Cloke, H. L., & Pappenberger, F. (2020). Evaluation of the consistency of ECMWF ensemble forecasts. *Geophysical Research Letters*, *46*, e2020GL087934. https://doi.org/10.1029/2020GL087934

Received 12 MAR 2020
Accepted 29 APR 2020
Accepted article online 3 MAY 2020

Abstract An expected benefit of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts. Inconsistent (jumpy) forecasts can cause users to lose confidence in the forecasting system. We present a first systematic, objective evaluation of the consistency of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble using a measure of forecast divergence that takes account of the full ensemble distribution. Focusing on forecasts of the North Atlantic Oscillation and European Blocking regimes up to 2 weeks ahead, we identify occasional large inconsistency between successive runs, with the largest jumps tending to occur at 7–9 days lead. However, care is needed in the interpretation of ensemble jumpiness. An apparent clear flip-flop in a single index may hide a more complex predictability issue which may be better understood by examining the ensemble evolution in phase space.

Plain Language Summary Ensemble forecasts show the range of weather scenarios that can occur, allowing users to make appropriate risk-based decisions. An ensemble forecast made 2 weeks in advance will show a range of possible outcomes. New observations included in subsequent forecasts will eliminate some of these scenarios, and the forecast will become more certain. Occasionally, a new forecast seems to contradict the previous forecast by introducing a new weather scenario that was not represented in the earlier forecast. Such inconsistencies can cause users to lose confidence in the forecasting system. We present a new method to assess the consistency of ensemble forecasts of large-scale weather patterns over Europe made by the European Centre for Medium-Range Weather Forecasts. We show that a careful analysis of each forecast is needed to understand how and why these jumps occur. Understanding and reducing the occurrence of inconsistent ensemble forecasts will increase user confidence and improve decision making.

1. Introduction

The chaotic nature of the atmosphere means that numerical weather prediction (NWP) forecasts are sensitive to small changes in their initial conditions. Operational NWP centers address this by running a number of forecasts from similar starting conditions. The resulting ensemble of forecasts shows the range of future atmospheric states consistent with the known uncertainties in the initial conditions (Leutbecher & Palmer, 2008; Swinbank et al., 2016). One of the expected benefits of ensemble forecasts is that a sequence of consecutive forecasts valid for the same time will be more consistent than an equivalent sequence of individual forecasts (Buizza, 2008; Zsoter et al., 2009). Inconsistent (or jumpy) forecasts are difficult to handle and can cause users to lose confidence in the forecasting system (Hewson, 2020; Pappenberger et al., 2011). However, this aspect of ensemble forecasts has received little attention in the literature.

The inconsistency between successive ensemble-mean (EM) forecasts valid for the same time was investigated by Zsoter et al. (2009). They define an inconsistency index as the difference between two fields over a given area, divided by their average standard deviation over the area. They consider cases of large jumps (inconsistency greater than a chosen threshold) and focus on sequences of jumps of opposite sign (flip-flops). Using this methodology, they showed that EM forecasts are more consistent than the corresponding ensemble control forecasts. Zsoter et al. (2009) conclude by noting that to further investigate the benefit of ensemble forecasts compared to single forecast, an index for probabilistic forecasts will need to be developed. Forecast consistency has also been considered in the context of model output statistics

© 2020 The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

(Ruth et al., 2009), comparing automated with manual forecasts (Griffiths et al., 2019), comparing deterministic rainfall forecasts from different models (Ehret, 2010) and in forecasts of river flow (Pappenberger et al., 2011).

None of the above methods are directly applicable to assess the consistency of a sequence of ensemble forecasts taking account of the full ensemble distribution. In this work, for the first time, we investigate the consistency of the European Centre for Medium-Range Forecasts (ECMWF) ensemble (ENS) using a measure of forecast divergence that accounts for all aspects of the ensemble empirical distribution.

We focus on two key characteristics of the large-scale flow over the European-Atlantic region: the North Atlantic Oscillation (NAO) and Scandinavian Blocking (BLO). Predicting transitions between such large-scale weather regimes 2 weeks or more ahead is a significant scientific challenge and at the frontier of NWP (ECMWF, 2015). These transitions are associated with large-scale changes in temperature and winds over Europe (Ferranti et al., 2018; Yiou & Nogaj, 2004) and hence have significant societal impacts, for example, on health (Charlton-Perez et al., 2019) and on energy production (Grams et al., 2017). We consider the full 15-day forecast range of the operational ENS.

The data and indices used are introduced in section 2. Methods, including the definition of the forecast divergence, are described in section 3. We then evaluate the inconsistency of the ENS forecasts for NAO and BLO and compare the jumpiness of the ENS with that of the EM and control forecasts in section 4. We present concluding remarks and avenues for future work in section 5.

2. Data

We study the time evolution of the NAO and BLO patterns that are associated with high-impact temperature anomalies over Europe (Ferranti et al., 2018). Following the approach of Ferranti et al. (2018), we use a two-dimensional phase space based on the two leading Empirical Orthogonal Functions (EOFs) of mid-tropospheric flow computed over the Euro-Atlantic region. The EOFs are computed using daily geopotential height at 500 hPa computed for the Euro-Atlantic region (30°N to 88.5°N, 80°W to 40°E) from 29 years of extended winter periods (October to March) of ECMWF ERA-Interim data (Berrisford et al., 2011; Dee et al., 2011). For the EOF computation, a 5-day running mean was used, and the mean seasonal cycle was removed. The first EOF represents the positive phase of the NAO (NAO+): a negative anomaly over Iceland and positive anomaly to the south (Cassou, 2008). The second EOF has a positive anomaly (high pressure) over Scandinavia, and a low to the east over the Atlantic, representing the flow pattern associated with blocking events over northern Europe (Ferranti et al., 2015). We refer to Ferranti et al. (2018) for further details.

We study the consistency of the operational ECMWF ensemble forecasts (ENS; Ben Bouallègue et al., 2019; Buizza & Richardson, 2017) of the large-scale flow over the North Atlantic Europe region for DJF 2016–2019, that is, 1 December 2015 to 28 February 2019, a total of 361 cases. All forecasts verifying at 00 UTC between 1 December and 28/29 February are included in the evaluation. The ENS comprises 50 perturbed members and one control member. The forecasts are valid for lead times of 1 to 15 days (at 24-hr intervals). The 500 hPa fields of each ENS forecast are extracted on a 1×1 degree grid and projected onto the two EOFs. The projections describe the magnitude of the NAO and BLO in each forecast, calculated relative to the climatological standard deviation. Following Ferranti et al. (2018), cases with projections greater than one standard deviation are considered large amplitude events.

3. Methods

We consider a sequence of ensemble forecasts valid for the same time t_v and started from initial conditions between 1 and L days before, $f(t_v, i), i = 1, \ldots L$. Each ensemble consists of M members, $f_m(t_v, i), m = 1, \ldots M$. We consider NAO and BLO separately, so f_m are univariate and real-valued.

To measure the difference between two ensembles f and g with M and N members, respectively, we use the divergence function given by

$$d(f, g) = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|f_i - g_j\right| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left|f_i - f_j\right| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|g_i - g_j\right|.$$

d is the divergence function associated with the Continuous Ranked Probability Score (CRPS), which is widely used to measure of the quality of ensemble forecasts (Gneiting & Raftery, 2007). If either M or N is equal to one, then d reduces to the CRPS, while if both are one, d is simply the absolute distance $|f - g|$. This means that d can also be used to measure the difference between two EM or control forecasts. d shares the important property of propriety with CRPS (Gneiting & Raftery, 2007), and as shown by Thorarinsdottir et al. (2013), these properties make d a particularly suitable choice.

The difference between two ensemble forecasts initialized on consecutive days and valid for the same time is

$$D(t_v, i) = d(f(t_v, i), f(t_v, i-1)),\ i = 1, \ldots, L,$$

where $f(t_v, 0)$ is the set of initial perturbed ensemble members at time t_v.

To measure the overall divergence (or inconsistency) between the sequence of forecasts valid for a given time, we sum the divergence between successive pairs of forecasts. To focus on the jumpiness within the sequence rather than a general trend across lead times (or a single large jump representing a one-time change in predictability), we subtract the difference between the first and last forecast of the sequence and define the divergence index (DI) for a given case as

$$\mathrm{DI}(t_v) = \frac{1}{L-1}\left(\left(\sum_{l=1}^{L} D(t_v, l)\right) - d(f(t_v, L), f(t_v, 0))\right).$$

The DI is calculated for the ENS and also for the ensemble control forecast (CTRL) and the EM. We refer to DI (ENS), DI (CTRL), and DI (EM), respectively. In this study, all ensemble forecasts have $M = 50$ members (control not included), and we consider forecasts up to lead time of $L = 15$ days.

As noted above, for a single forecast such as CTRL and EM, the divergence is equal to the absolute difference. For these forecasts, DI is similar (though not identical) to the flip-flop index of Griffiths et al. (2019).

4. Results

Figure 1 (upper panel) shows the DI (ENS) for NAO (solid) and BLO (dashed) for each day of the last four winters (December–February 2015–2016 to 2018–2019; the vertical dotted lines indicate the start of each season). Positive values indicate higher inconsistency. There is similar variability in DI for both regimes (standard deviation of 0.027 for NAO and 0.028 for BLO). However, the peaks of high/low consistency occur at different times. Winter 2018–2019 was more inconsistent than usual for blocking, while forecasts for NAO were not unusually inconsistent in this season. Overall, there is no strong correlation between inconsistency in forecasts of blocking and NAO (correlation = −0.1 over the full set of cases).

To illustrate the different levels of consistency associated with the high and low DI, three example cases are shown in the lower panel (labeled A, B, and C on top panel). B (center) shows an example of a case with very good consistency in forecasting the BLO regime. The plot shows the amplitude of blocking for 14 December 2017 predicted by forecasts initialized between 30 November and 13 December. The 15-day ENS forecast has a broad distribution (large spread), similar to the climate distribution. Subsequent forecasts show smaller spread and a consistent shift of the ENS towards negative BLO.

C (right) shows a contrasting case with poor consistency in forecasting blocking. The plot shows the amplitude of blocking for 14 December 2018 predicted by forecasts initialized on 30 November to 13 December. The longest range forecasts are similar to the climate distribution; there is a trend over the following days showing an increasing probability for blocking. However, there is then an abrupt change in the forecast to a strong signal for neutral conditions, followed by an equally abrupt change back to blocking. This is the most inconsistent BLO case of this whole period.

A (left) is a case of large inconsistency for the NAO. This occurs at the end of an extended period of strong NAO− (and associated cold weather over NW Europe). The forecasting challenge in this case is to identify

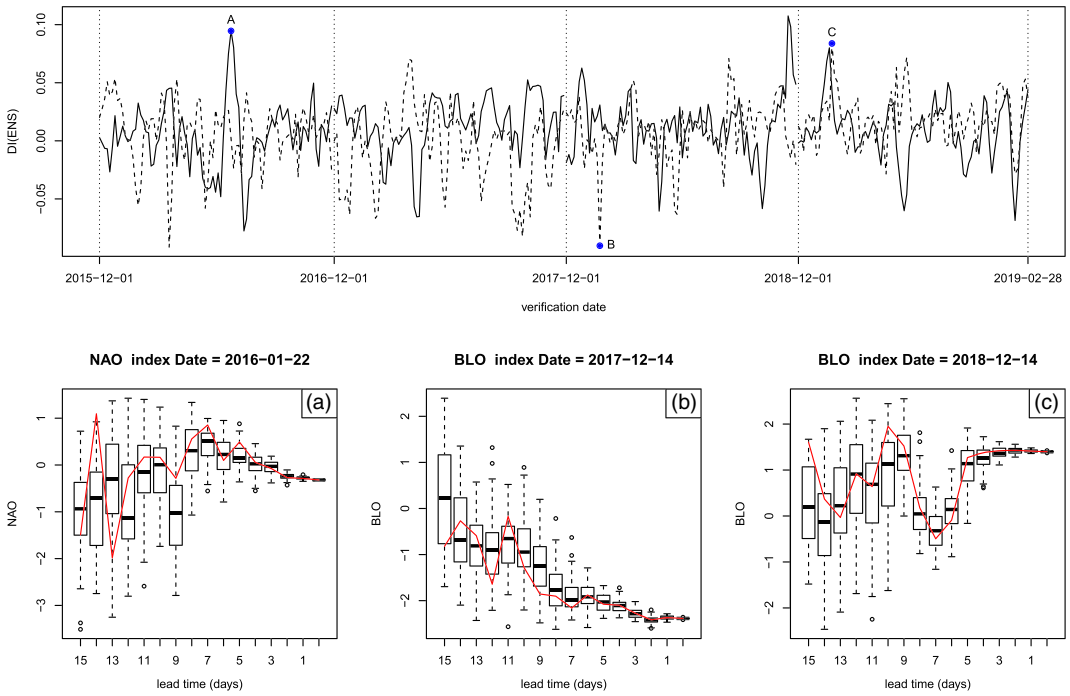

Figure 1. Consistency in ENS forecasts of the NAO and BLO regimes. Upper panel: time series of overall consistency DI (ENS) of 1–15 day forecasts verifying during winters (December–February) 2015–2019. Positive values indicate lower consistency. Lower panels show examples of both consistent and inconsistent cases for each regime. Each example shows the distribution of ENS forecasts verifying for a given date with lead time of 1 to 15 days; box and whiskers show min, max, and 25, 50, and 75 percentiles of the ENS distribution (50 perturbed members); red line shows the ENS control.

when this cold event will end. The longest range forecasts show large uncertainty but with probability of around 50% for a return to near-normal conditions (NAO magnitude <1). The forecasts from 11 January onwards show much higher probability for the end of the NAO− event, with the exception of the forecast from 13 January which again gives a higher probability for the cold spell to continue beyond 21 January.

These cases of large inconsistency illustrate the challenge for users—in both, there is an apparent increase in certainty for a change in weather type (regime). But this is thrown into doubt by a large change in a subsequent forecast. The following jump back is also difficult for the user to manage—can it be trusted, or will the following forecast jump again? While such cases are uncommon in the ENS (Figure 1, top), they nevertheless can cause a loss of confidence in the forecasts and merit further investigation.

The consistency of ENS is compared with that of the control forecast and of the EM in Figure 2 for NAO (results for BLO are similar). Overall, DI is much larger for the EM (mean DI 0.14) and especially CTRL (0.42) than for ENS (0.01), reflecting how the full ENS distribution does mitigate the jumpiness seen in the deterministic forecasts. The cases with large DI (ENS) also tend to have large DI (EM), and vice versa. The examples of inconsistent ENS forecasts in Figure 1 are typical—there is a substantial shift of the whole ENS distribution, which is reflected in both DI (EM) and DI (ENS). For more consistent cases, the correlation is less strong. When the whole ENS distribution is very consistent, the EM must also be consistent. However, when the EM is consistent, there may still be variation in the ENS distribution as a whole (for example, changes in spread) that can lead to larger DI (ENS).

There is much less correlation between DI (ENS) and DI (CTRL). The most inconsistent cases for ENS tend to be associated with a substantial shift in the whole ENS distribution, and the control also shows large inconsistency as expected. However, there are also cases with large DI (CTRL) but small DI (ENS)—large jumps in CTRL are not reflected in the ENS as a whole, as seen in the examples. This is an important result that demonstrates that jumpiness in the ENS is not simply a consequence of a corresponding jumpiness in the CTRL.

Figure 3 shows the distribution of magnitude of the individual jumps ($|D(t_v, i)|$, absolute value of difference between forecasts started 1 day apart) at each lead time for both ENS and CTRL. The two inconsistent cases

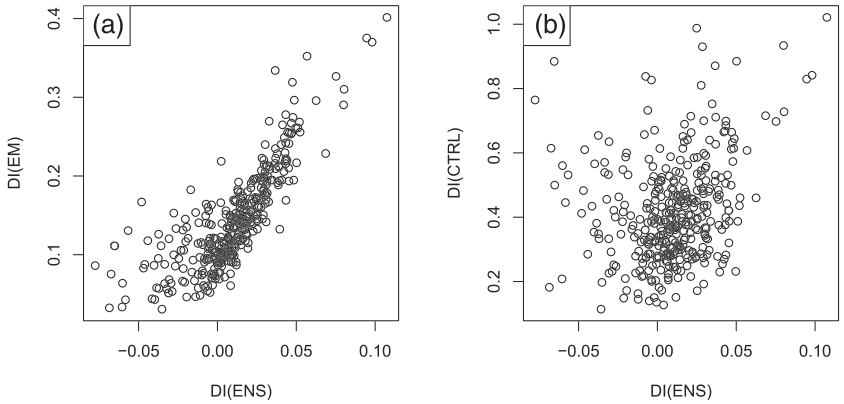

Figure 2. Comparison of consistency of ENS, CTRL, and EM for NAO. Each panel shows a scatter plot of DI (ENS) on the *x*-axis against (a) DI (EM) and (b) DI (CTRL); 361 cases verifying during winters (December–February) 2015–2019.

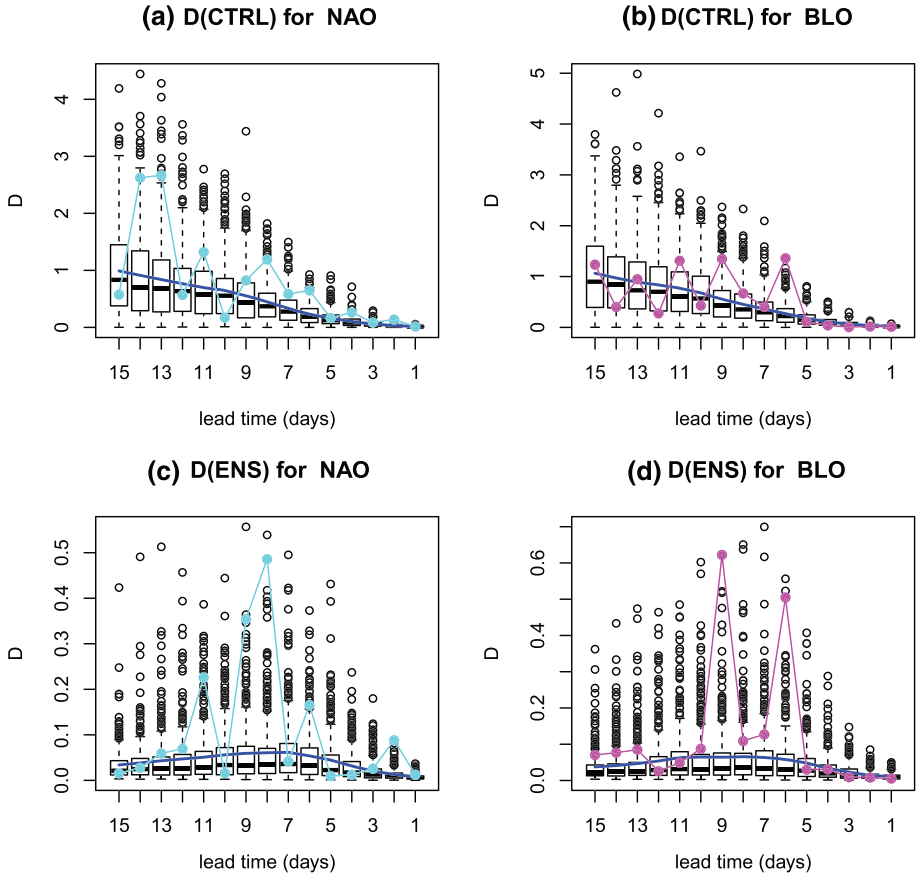

Figure 3. Distribution of jumps ($|D(t_v, i)|$) at each forecast lead time (i days) for CTRL (top) and ENS (bottom) for the NAO (left) and BLO (right) regimes. Box and whiskers show 25, 50, and 75 percentiles of the ENS distribution, with outliers shown by open circles; thick blue lines show the mean value. The values for the sequence of forecasts verifying on 22 January 2016 for NAO (cyan) and 14 December 2018 (magenta) correspond to the two examples of inconsistent forecasts shown in Figure 1.

A and C from Figure 1 are highlighted. As well as having large overall DI (ENS), both cases have some of the largest individual ENS jumps between consecutive forecasts at any lead time. As for DI, the magnitude of the individual jumps is much larger for CTRL than for ENS.

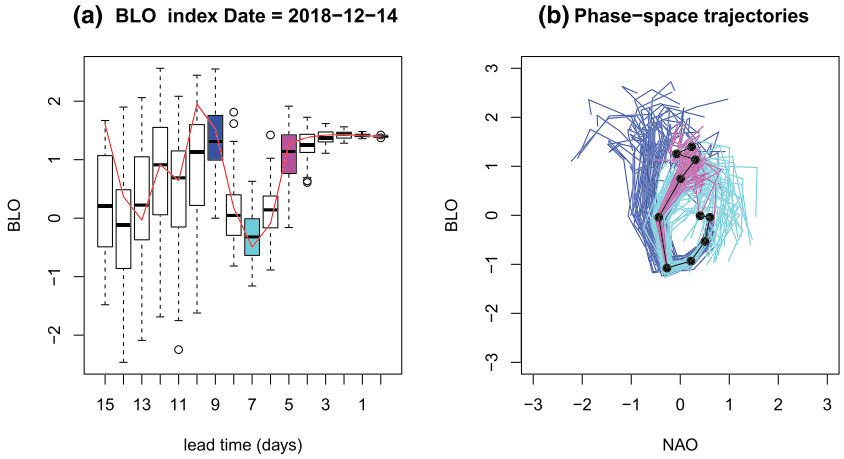

Figure 4. Phase-space trajectories of ENS forecasts initialized on 5, 7, and 9 December 2018. (a) Amplitude of blocking for 14 December 2018 predicted by forecasts from different initial times up to 15 days ahead with forecasts from 5, 7, and 9 December highlighted; box and whiskers show min, max, and 25, 50, and 75 percentiles of the ENS distribution, with outliers shown by open circles; red line shows the ENS control. (b) Phase-space trajectories of ENS forecasts initialized on 5, 7, and 9 December 2018 (blue, cyan, and magenta, respectively) and verifying analysis trajectory (black).

Figure 3 highlights another important difference between the jumpiness of the ENS and CTRL. For CTRL, $|D(t_v, i)|$ increases with lead time, with the mean jump approaching 1 by day 15. However, for ENS, the largest mean value and most extreme jumps tend to occur at around 7–9 days lead. At longer lead times, as memory of the initial conditions is lost, the limit of predictability is reached and each forecast behaves like a random draw from the climate distribution. This means that at long lead, the difference between two control forecasts will be on average the same as the difference between two randomly selected states from the climate (see Text S1 in the supporting information for details). In contrast, at this range, two ENS forecasts will represent two statistically indistinguishable samples from the same climate distribution. Any difference between them will only be due to sampling, and for a sufficiently large ensemble, $D(t_v, i)$ will be small.

We have seen that DI can identify cases of high inconsistency in the ENS. A more detailed investigation of such cases is merited to understand what aspects of the ensemble forecast configuration lead to such behavior. The high-DI cases, A and C (Figure 1) both occur in situations of transitions between large-scale regimes. A compact way to visualize these transitions is in a phase-space plot which can be used to examine how the magnitude of both BLO and NAO evolve through the forecast for each ensemble member (Ferranti et al., 2018). Following this approach for high-DI cases also brings some new insight into the jumpiness itself.

To illustrate this, we consider the BLO case of 14 December 2018 (C in Figure 1) and examine the phase-space trajectories of the relevant forecasts. We compare the forecasts started on 5 and 9 December (which both predict a positive BLO pattern) with the contrasting forecast from 7 December which has largest probability for a negative BLO to occur (Figure 4a). Figure 4b (and Figure S1) shows the phase-space evolution of the forecasts from 5, 7, and 9 December 2018. The forecast from 9 December follows the observed trajectory with only a few members moving too quickly away from the block. The forecast from 7 December also follows the observed trajectory for the first 4–5 days of the forecast, but then most members fail to maintain the blocking and evolve too quickly towards the more mobile NAO+ pattern, leading to the poor 7-day forecast for BLO (Figure 4a, cyan). The forecast from 5 December does not follow the observed trajectory so well from 9 December onwards: most ENS members move too quickly into a strong blocking and NAO−. Although this forecast gives a strong indication of blocking for 14 December (day 9 forecast, Figure 4a, blue), the evolution leading to this is clearly inconsistent with the observed development. While Figure 4a suggests that the forecast from 7 December has lost the signal that was present in earlier forecasts, the analysis of the phase-space trajectories shows that the situation was more complex. In fact, the forecast from 7 December better captured

the observed evolution up to 11 December, with significantly smaller ENS spread. Neither the 5 December nor the 7 December forecast captured the observed trajectory after this time. It was only the later forecasts, from 9 December onwards that correctly predicted the observed evolution.

This shows us that care is needed in the interpretation of the ensemble jumpiness. An apparent clear flip-flop in a single index may hide a more complex predictability issue. When investigating the cause of a case of high DI, it is important to frame the analysis in the right context, as shown by Figure 4. From a diagnostic point of view, Figure 4a raises the question: why do the forecasts from 7 December lose the signal that was present in the earlier forecast from 5 December? In contrast, looking at the wider context of Figure 4b raises the question: what mechanism caused the two successive changes in predictability, first to avoid the too strong NAO−/BLO (5 December forecast) and second to maintain the block and not move too quickly to NAO+ (7 December forecast). Error tracking (Grams et al., 2018; Magnusson, 2017) shows that both these errors can be traced back to the initial mishandling of developing trough-ridge patterns over eastern North America (Figures S2 and S3).

5. Conclusions

Predicting transitions between large-scale weather regimes 2 weeks ahead is a significant forecasting challenge. Occasionally, successive ensemble forecasts can give contradictory indications about the probability for a change in weather type. Such jumpiness or "flip-flopping" is difficult for users to manage since the forecast does not give a consistent message for decision making. While such cases are uncommon (Figure 1), they nevertheless can cause a loss of confidence in the forecasts and merit further investigation.

For the first time, we have carried out a systematic, objective evaluation of the consistency of ECMWF ensemble forecasts that takes account of the full ensemble distribution. This extends the earlier work of Zsoter et al., 2009 who focused specifically on flip-flops of the EM.

We investigated the ENS consistency for two key flow patterns for Europe, NAO and blocking. We used a measure of the divergence between two ensembles started at different times but valid for the same time. This allowed us to quantify both individual jumps and the overall consistency of a sequence of ENS forecasts valid for a given time. Our main conclusions are the following:

- In general, the peaks of high and low consistency occur at different times for NAO and BLO; there is no strong correlation between inconsistency for NAO and BLO (Figure 1).
- DI for the ENS is on average much lower than for EM and especially for CTRL (Figure 2) demonstrating benefit of the ensemble in mitigating the jumpiness of the deterministic forecasts by representing the range of possible scenarios.
- The largest individual jumps for ENS tend to be days 7–9, while for the CTRL the magnitude of individual jumps continues to increase throughout the forecast (Figure 3). This is associated with the different asymptotic behavior of the (deterministic) CTRL forecast and the ENS at long forecast lead.
- Care is needed in the interpretation of the ensemble jumpiness. What looks at first sight to be a clear case of flip-flopping in a single index (BLO or NAO) may be a more complex predictability issue. This may be better understood by examining the phase-space evolution of both components together (Figure 4).

In this work, we assessed the consistency of the univariate forecast of NAO and BLO separately. However, we also showed how it is important to consider the ensemble trajectories in the two-dimensional phase to properly understand the reason for apparent jumpiness. It will therefore be valuable to extend the divergence and DI methodology to the multivariate situation so that the consistency of NAO and BLO can be evaluated together. This will also enable investigation of the consistency of other aspects of ensemble performance such as for tropical cyclone tracks.

The DI allows us to identify important cases of high ensemble forecast inconsistency and to routinely monitor the occurrence of such cases. Careful diagnosis of these cases will help to identify the causes of the inconsistency and hence to address the relevant aspects of ensemble configuration and modeling. Reducing the occurrence of inconsistent (or jumpy) ensemble forecasts will increase user confidence and improve decision making.

Acknowledgments

David Richardson is supported by a Wilkie Calvert PhD Studentship at the University of Reading. The data used for the EOF computations is available from the ECMWF ERA-Interim archive (Berrisford et al., 2011): https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim. The ECMWF ensemble forecast data used in this research is available on the TIGGE archive (Swinbank et al., 2016): https://confluence.ecmwf.int/display/TIGGE.

References

Ben Bouallègue, Z., Magnusson, L., Haiden, T., & Richardson, D. S. (2019). Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events. *Quarterly Journal of the Royal Meteorological Society*, *145*(721), 1741–1755. https://doi.org/10.1002/qj.3523

Berrisford, P., Dee, D. P., Poli, P., Brugge, R., Fielding, M., Fuentes, M., et al. (2011). The ERA-Interim archive Version 2.0. ERA Report Series, (1), 23. Retrieved from https://www.ecmwf.int/node/8174.

Buizza, R. (2008). The value of probabilistic prediction. *Atmospheric Science Letters*, *9*(2), 36–42. https://doi.org/10.1002/asl.170

Buizza, R, & Richardson, D. (2017). 25 years of ensemble forecasting at ECMWF|ECMWF (No. 153). ECMWF Newsletter. https://doi.org/10.21957/bv418o

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

Charlton-Perez, A. J., Aldridge, R. W., Grams, C. M., & Lee, R. (2019). Winter pressures on the UK health system dominated by the Greenland Blocking weather regime. *Weather and Climate Extremes*, *25*, 100218. https://doi.org/10.1016/j.wace.2019.100218

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

ECMWF. (2015). The strength of a common goal. Strategy 2016-2025. Retrieved from https://www.ecmwf.int/sites/default/files/ECMWF_Strategy_2016-2025.pdf.

Ehret, U. (2010). Convergence index: A new performance measure for the temporal stability of operational rainfall forecasts. *Meteorologische Zeitschrift*, *19*(5), 441–451. https://doi.org/10.1127/0941-2948/2010/0480

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.

Ferranti, L., Magnusson, L., Vitart, F., & Richardson, D. S. (2018). How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? *Quarterly Journal of the Royal Meteorological Society*, *144*(715), 1788–1802. https://doi.org/10.1002/qj.3341

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Grams, C. M., Beerli, R., Pfenninger, S., Staffell, I., & Wernli, H. (2017). Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, *7*(8), 557–562. https://doi.org/10.1038/nclimate3338

Grams, C. M., Magnusson, L., & Madonna, E. (2018). An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2577–2591. https://doi.org/10.1002/qj.3353

Griffiths, D., Foley, M., Ioannou, I., & Leeuwenburg, T. (2019). Flip-flop index: Quantifying revision stability for fixed-event forecasts. *Meteorological Applications*, *26*(1), 30–35. https://doi.org/10.1002/met.1732

Hewson, T. (2020). Use and verification of ECMWF products in member and co-operating states (2019). *ECMWF Technical Memorandum*, (860). https://doi.org/10.21957/80s471ib1

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Magnusson, L. (2017). Diagnostic methods for understanding the origin of forecast errors. *Quarterly Journal of the Royal Meteorological Society*, *143*(706), 2129–2142. https://doi.org/10.1002/qj.3072

Pappenberger, F., Bogner, K., Wetterhall, F., He, Y., Cloke, H. L., & Thielen, J. (2011). Forecast convergence score: A forecaster's approach to analysing hydro-meteorological forecast systems. *Advances in Geosciences*, *29*, 27–32. https://doi.org/10.5194/adgeo-29-27-2011

Pappenberger, F., Cloke, H. L., Persson, A., & Demeritt, D. (2011). HESS opinions "on forecast (in)consistency in a hydro-meteorological chain: Curse or blessing?". *Hydrology and Earth System Sciences*, *15*(7), 2391–2400. https://doi.org/10.5194/hess-15-2391-2011

Ruth, D. P., Glahn, B., Dagostaro, V., Gilbert, K., Ruth, D. P., & Glahn, B. (2009). The performance of MOS in the digital age. *Weather and Forecasting*, *24*(2), 504–519. https://doi.org/10.1175/2008WAF2222158.1

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., et al. (2016). The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, *97*(1), 49–67. https://doi.org/10.1175/BAMS-D-13-00191.1

Thorarinsdottir, T. L., Gneiting, T., & Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 522–534. https://doi.org/10.1137/130907550

Yiou, P., & Nogaj, M. (2004). Extreme climatic events and weather regimes over the North Atlantic: When and where? *Geophysical Research Letters*, *31*, L07202. https://doi.org/10.1029/2003GL019119

Zsoter, E., Buizza, R., & Richardson, D. (2009). "Jumpiness" of the ECMWF and Met Office EPS control and ensemble-mean forecasts. *Monthly Weather Review*, *137*(11), 3823–3836. https://doi.org/10.1175/2009MWR2960.1