Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Trajectory adjustment of lagged seasonal forecast ensembles

Note no SAMBA/19/20
Authors Thordis L. Thorarinsdottir
Nina Schuhen
Alex Lenkoski

Date 10th June 2020

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

The authors

Thordis L. Thorarinsdottir and Alex Lenkoski are Chief Research Scientists at the Norwegian Computing Center, Nina Schuhen is Researcher at CICERO Center for International Climate Research

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Trajectory adjustment of lagged seasonal forecast ensembles**
Authors	**Thordis L. Thorarinsdottir , Nina Schuhen , Alex Lenkoski**
Date	10th June 2020
Publication number	SAMBA/19/20

Abstract

Seasonal forecasting has became a critical area of development in numerical weather prediction. Reliable forecasts beyond the two week time period are necessary for a number of industrial and societal planning applications and new approaches are being developed to extend the useful range of numerical weather prediction output. We investigate the performance of one such system, the UK Met Office's GloSea5 system, an ensemble system with the novel feature that ensemble members are initiated in a rolling and staggered manner. Focusing on summer surface temperatures, we show that individual model runs from this system do not exhibit skill beyond the two-week time horizon and indeed substantially under-perform climatological forecasts at longer lead times. However, when combining the ensemble system and applying the Rapid Adjustment of Forecast Trajectories (RAFT) methodology to the individual runs, we show that the combined forecast can achieve performance which is always at least on par with climatology and in many circumstances exhibits modest outperformance.

Keywords	Seasonal forecasting; Lagged ensemble; Postprocessing; Rapid adjustment of forecast trajectories (RAFT)
Target group	Scientists
Availability	Open
Project	PostProcessingPhD
Project number	220783
Research field	Statistics; Meteorology
Number of pages	13

1 Introduction

Weather forecasting beyond the medium range of two weeks is currently an active area of research (Robertson and Vitart, 2018) due to the demand for skillful long-range forecasts in various societal sectors such as energy production, agriculture, health and disaster management (e.g. Ogallo et al., 2008). Sources of long-range predictability within the atmosphere are usually associated with the existence of different modes of low-frequency variability, including the El Niño Southern Oscillation (ENSO), monsoon rains, sudden stratospheric warmings, the Madden Julian Oscillation (MJO), the Indian Ocean dipole, the North Atlantic Oscillation (NAO), and the Pacific/North American (PNA) pattern, spanning a wide range of time scales from months to decades (Hoskins, 2013; Vitart et al., 2012). It is expected that, if a forecasting system is capable of reproducing phenomena with low-frequency variability, they may also be able to forecast them (Van Schaeybroeck and Vannitsem, 2018). Post-processing and skill assessment of long-range forecasts is thus often focused on these same phenomena (e.g. Van Schaeybroeck and Vannitsem, 2018), or other slowly-evolving components of the Earth system such as sea-surface temperature (e.g. Heinrich et al., 2019). However, forecast users commonly need information on atmospheric variables such as surface temperature and precipitation (Roulin and Vannitsem, 2019).

At time scales beyond the medium range, the weather noise that arises from the growth of the initial uncertainty, becomes large (Royer, 1993). As a consequence, predictions must be probabilistic in nature. This is made possible through the use of ensemble forecasts (Van Schaeybroeck and Vannitsem, 2018), with a trade-off between increased computational costs and increased skill as the ensemble size grows. For monthly to seasonal forecasts, the benefit of good initialization (initialization as close as possible to observations) has been demonstrated (Doblas-Reyes et al., 2013a,b). For these reasons, the UK Met Office's seasonal prediction system, GloSea5, uses a lagged initialization approach with new ensemble members initialized every day, resulting in a monthly seasonal forecast ensemble with 42 members generated by combining all forecasts available from the most recent three weeks (MacLachlan et al., 2015)[1].

In this paper, we investigate how the older members of a lagged ensemble system can be brought closer to observations by utilizing new observations that have become available since the forecast system was run to generate these members, using the rapid adjustment of forecast trajectories (RAFT) algorithm recently proposed by Schuhen et al. (2020). With a focus on weekly average surface temperature, we aim to assess the skill of the forecast in a user-relevant setting. For observations, we use the ERA5 reanalysis. Preliminary data for ERA5 is now being released daily with a 5-day delay from real time, making the setting considered here somewhat realistic from an operational perspective. The data sets and the RAFT algorithm are described in the following Section 2, with results shown in Section 3. Finally, some concluding remarks are given in Section 4.

1. https://www.metoffice.gov.uk/research/climate/seasonal-to-decadal/gpc-outlooks/user-guide/technical-glosea5

2 Data and methods

2.1 Data

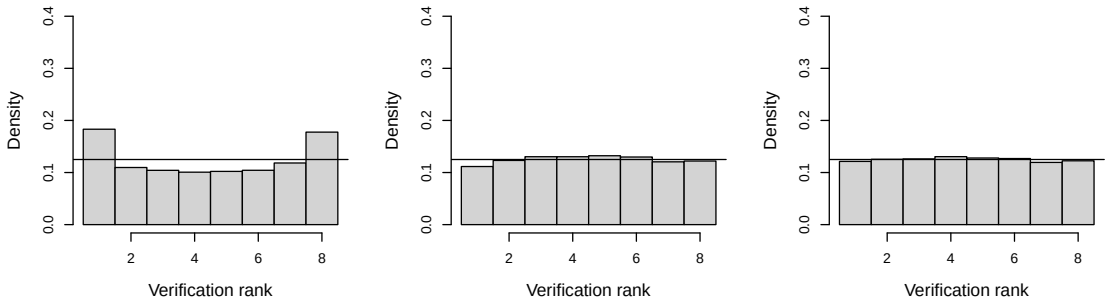

Figure 1. Verification rank histograms for GloSea5 forecasts of weekly mean temperature anomalies initialized on May 1st compared against ERA5. The results are aggregated over the study region, the time period 1993-2015 as well as lead times 1-6 weeks (left), 7-12 weeks (middle) or 13-18 weeks (right). The black horizontal lines indicate a perfectly calibrated forecast.

We analyze surface temperature hindcasts, or historical re-forecasts, from GloSea5, the UK Met Office Global Seasonal forecast system version 5 (MacLachlan et al., 2015). The GloSea5 system has a spatial resolution of 0.8 degrees in latitude and 0.5 degrees in longitude. Our analysis focuses on land grid cells in a region bounded by -30 to 50 longitude and 30 to 90 latitude, covering Europe and surrounding area. The hindcasts cover the time period 1993 to 2015, and the system uses a lagged initialization approach with seven members initialized on the 1st, 9th, 17th and 25th of every month. Hindcasts of weekly mean temperatures from five initialization dates–May 1st to June 1st–are considered for realization dates of up to 18 weeks ahead for the May 1st run, or the time period from early May to early September. The analysis is performed on temperature anomalies which are defined relative to the model's weekly climatology over the entire time period 1993-2015. In the remainder of the paper, we will refer to the hindcasts as "forecasts".

The GloSea5 forecasts are compared against the ERA5 reanalysis (Copernicus Climate Change Service (C3S), 2017). ERA5 originally has a spatial resolution of 0.28 degrees and is here upscaled to match the resolution of the GloSea5 system. We calculate weekly mean anomalies in the same manner as for the hindcasts using ERA5's climatology over the same time period.

The aim of the forecast system is to provide accurate and calibrated forecasts (e.g. Thorarinsdottir and Schuhen, 2018). Calibration, or reliability, refers to the representation of uncertainty in the forecast in that an event predicted to occur with probability p should be realized with the same frequency in the reanalysis. An empirical calibration assessment of the seven member ensemble initialized on May 1st is shown in Figure 1. The plots show the distribution of the rank of the reanalysis when compared against the seven ensemble members across years, spatial locations and forecast lead times. While the forecasts are slightly underdispersive for the first six weeks as indicated by the $\cup$-shape, they

are nearly perfectly calibrated for weeks 7-18. For this reason, we will in the following focus on improving the prediction accuracy.

2.2 Rapid adjustment of forecast trajectories (RAFT)

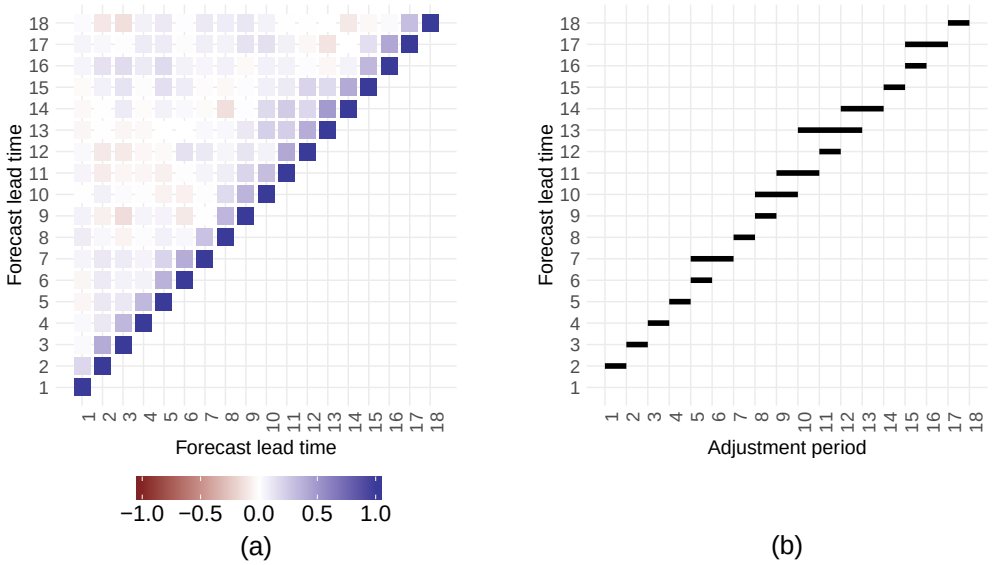

Figure 2. (a) Correlations between forecast anomaly errors at different lead times of the same forecast trajectory for the ensemble mean forecast initialized on May 1st; (b) The resulting adjustment periods for each forecast lead time.

To improve the accuracy of the forecasts, we consider new information that has become available since the forecast was issued, namely observations associated with lead times that have already been realized. Specifically, if the forecast errors at subsequent lead times are correlated with the most recently observed forecast error, this information can be used to update the remaining forecast trajectory that is yet to be realized using the rapid adjustment of forecast trajectories (RAFT) algorithm proposed by Schuhen et al. (2020). The forecast error $e_{t,l}$ is here defined as the distance of the ensemble mean anomaly forecast $\bar{x}_{t,l}$ initialized at time t and valid at lead time l to the observed anomaly y_{t+l} at time $t+l$,

$$e_{t,l} = y_{t+l} - \bar{x}_{t,l}. \tag{1}$$

Figure 2(a) shows the correlation between forecast errors at different lead times for the ensemble mean forecast trajectory initialized on May 1st. While the errors at all lead times beyond the first show substantial correlation with the error observed at the previous lead time, the correlation decreases rapidly for lead times further into the future.

We use a linear regression model to connect the error at a future lead time $l' > l$ with the current error $e_{t,l}$. Specifically, we define the model

$$e_{t,l'} = \alpha + \beta\, e_{t,l} + \varepsilon, \tag{2}$$

where α and β are real valued regression coefficients and ε is a normally distributed error term with mean zero. The model is estimated separately for each forecast run, current lead time and future lead time in a leave-one-out cross-validation approach, i.e. forecast anomaly errors for each year are predicted by using data from all remaining years. In

Schuhen et al. (2020) and Schuhen (2019), the number of future lead times that are corrected each time is selected based on a hypothesis test for $\beta = 0$ after estimating the regression equation in (2) for future lead times $l+1, l+2, \ldots$. At the first future lead time l^* where this test is not rejected, the procedure is stopped and only lead times l' with $l < l' < l^*$ are corrected. Here, this approach turns out to produce unrealistically long adjustment periods and thus spurious correlations can result in reduction of the forecast accuracy rather than an improvement. As we only have a small number of lead times, we instead determine the length of the adjustment periods empirically.

For each l' with $l < l' < l^*$, we then update the ensemble mean forecast $\bar{x}_{t,l'}$ to $\bar{x}_{t,l'} + \hat{e}_{t,l'}$ where $\hat{e}_{t,l'}$ is the estimated error based on (2). The adjustment periods for the forecast run initialized on May 1st are shown in Figure 2(b). Further details of the RAFT algorithm are given in Schuhen et al. (2020) and Schuhen (2019).

3 Results

For evaluating the forecasts, we calculate the root mean square error (RMSE) skill score of the ensemble mean forecast with the ERA5 climatology forecast of that week and grid cell over the entire time period 1993-2015 as a reference forecast. A positive skill score indicates a higher skill than the climatology, while a negative skill score indicates a lower skill.

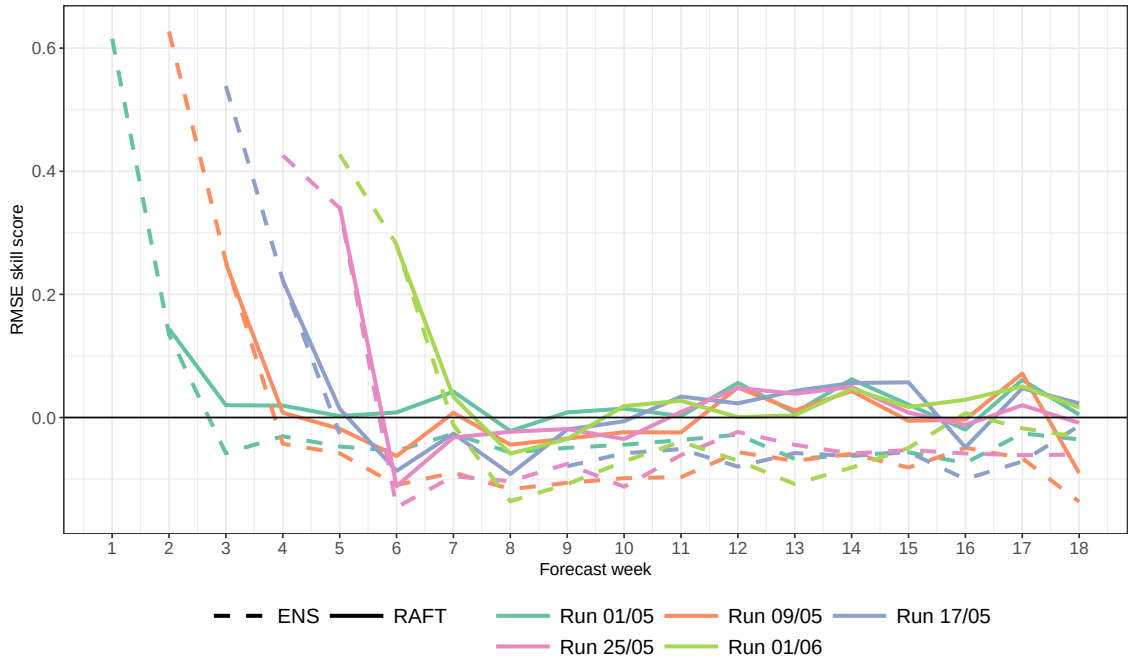

Figure 3. Root mean squared error (RMSE) skill scores for five different runs of GloSea5 compared against ERA5 climatology. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. The original GloSea5 forecasts are indicated with dashed lines while RAFT forecasts updated one week prior to the realization time are indicated by solid lines.

Figure 3 shows the RMSE skill scores for each of the GloSea5 runs as a function of lead time, aggregated over grid cell locations and years. All the runs show a similar pattern: In the first week, the forecast improves the climatological reference forecast by 40-60%, and in the second week, the forecasts are 15-35% better than climatology. From week three and onward, however, the skill is roughly constant at 5-15% below climatology. At the shortest possible adjustment lead time of one week, the forecasts updated with the RAFT algorithm are consistently better than the original forecasts and, on average, more skillful than the climatology forecast. The skill of RAFT forecasts with adjustment lead times from one week to that of the original forecast generally falls between the two forecasts shown in Figure 3. For example, for the run initialized on May 1st, Figure 2(b) shows that the adjustment period for this run varies from one to three weeks depending on the week. At any given time, the RAFT forecast trajectory will thus converge to the original forecast trajectory after one to three weeks. Results for the other four runs are similar (results not shown). On their own, the individual runs thus do not provide forecasts of higher skill than climatology beyond the medium range of two weeks.

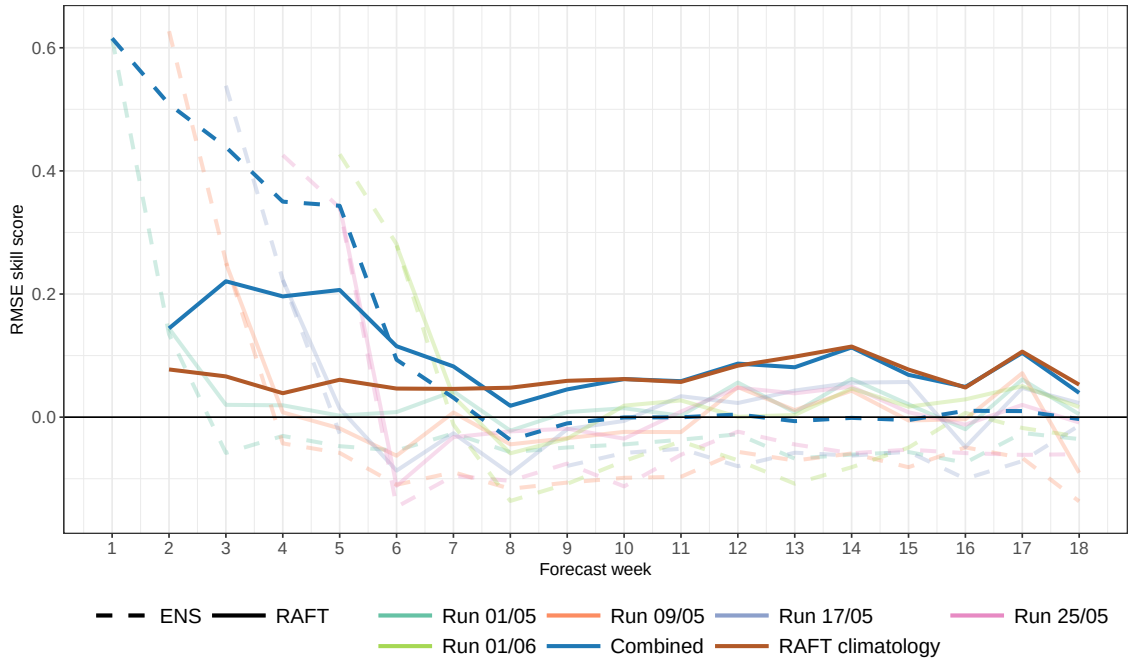

Figure 4. Root mean squared error (RMSE) skill scores for a comparison against ERA5 climatology for the combination of all five forecast runs (blue dashed line), for RAFT-processed ERA5 climatology of adjustment lead time one week (brown solid line), and for a combination of the RAFT-processed ensemble means for all five forecast runs at an adjustment lead time of one week (blue solid line). For comparison, these skill scores are overlaid on the results shown in Figure 3.

We now consider various forecast combinations where, in each case, the multi-model or lagged ensemble mean forecast is constructed using equal weights on the different models. As shown in Figure 4, for the first five forecast weeks, the skill of the lagged ensemble mean is slightly below that of the newest forecast run. From forecast week six and onward, no new runs are added to the lagged ensemble mean, resulting in increasing

effective lead time of the forecast. While this gradually reduces the skill, as expected, the reduction halts at around the skill of the climatology and beyond week eight, the two forecasts are comparable in skill. Thus, while the skill of each individual run is lower than that of climatology beyond the medium range of two weeks, their joint skill is consistently higher for lead times of up to three weeks and comparable thereafter.

The climatological reference forecast may be updated in the same manner as the GloSea5 forecasts using the RAFT algorithm. This results in a climatological forecast with a structure comparable to an autoregressive process of order one. The updated climatological forecast with lead time of one week is indicated with a brown line in Figure 4. This forecast has 5-10% higher skill than the climatological reference forecast. Furthermore, a forecast that combines the lagged ensemble means post-processed with RAFT is the best forecast for weeks 6 and 7, and comparable to the RAFT climatology for week 8 and onward.

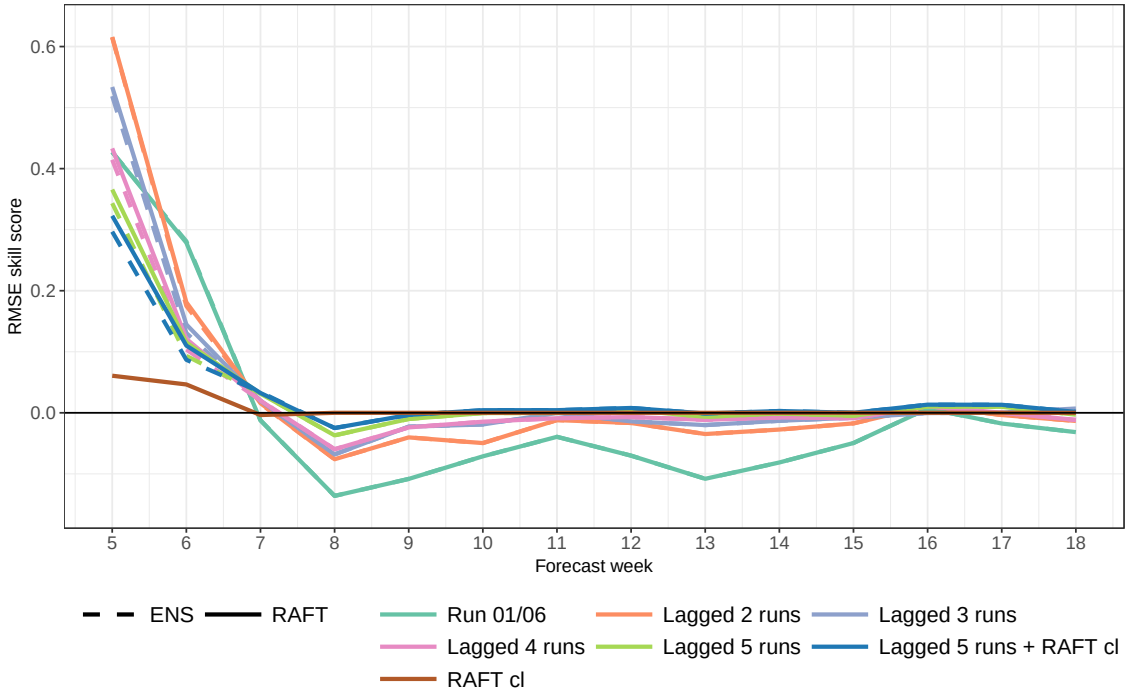

Figure 5. Root mean squared error (RMSE) skill scores for various model combinations for forecasts issued in week five compared against ERA5 climatology. Each run combination consists of the most recently available runs. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015.

For a further comparison of various model combinations in an operational setting, Figure 5 shows the skill scores for a number of forecasts for weeks 5-18 issued in week 5. These results indicate that an optimal forecasting strategy is to combine a smaller number of the most recent runs for the first two forecast weeks after which all five runs as well as the climatology should be combined. While the combination of all runs and climatology does not outperform climatology for all forecasts weeks, it is overall the best forecast for weeks 7-18. In particular, including climatology in the ensemble is consistently slightly

better than only considering the five GloSea5 runs.

As shown in Figure 2, the adjustment period at forecast lead time 5 is relatively short. This can also be seen in Figure 5 where the RAFT-adjusted forecasts coincide with the original forecasts from week 7. Figure 6 shows the same original forecasts from week 10 and onward, as well as the RAFT-adjusted forecasts issued in week 10. Here, the adjustment periods are considerably longer for all the forecast runs and the RAFT adjustment yields improved performance until week 14 after which the forecasts again coincide. In this case, the original forecasts have lead times of 5+ weeks, and we see that only the combination of all five runs is on a par with climatology. For a combination of three or more runs, the RAFT adjustment yields an overall higher skill than climatology with the full combination of all five runs and RAFT climatology again showing the highest skill overall.

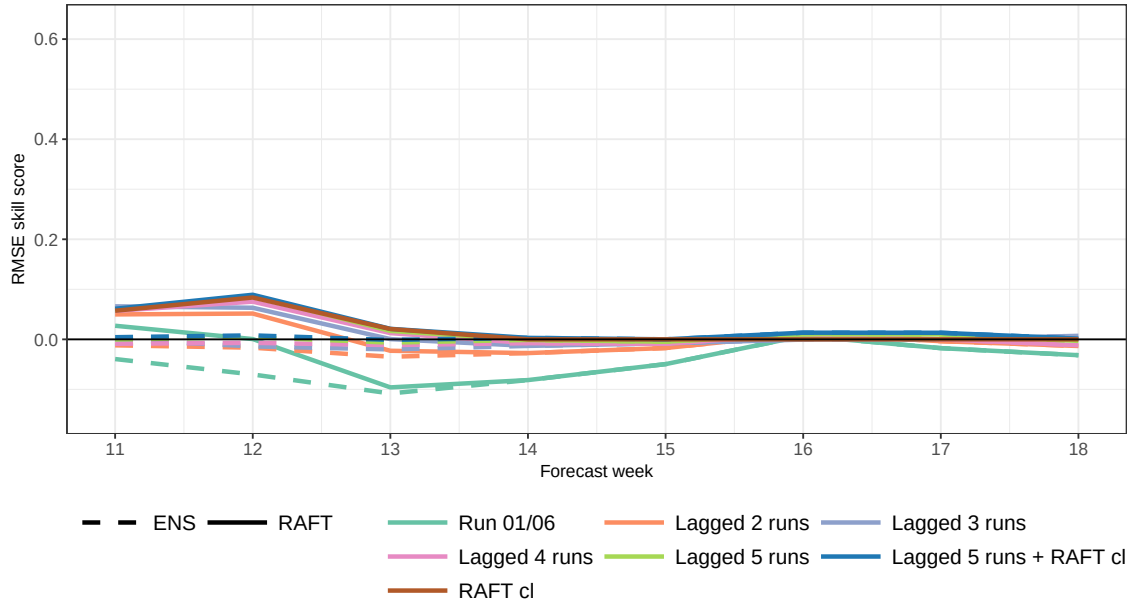

Figure 6. Root mean squared error (RMSE) skill scores for various model combinations compared against ERA5 climatology for forecast weeks 10-18. The score for each forecast week is aggregated over all land grid cells in the study area and the years 1993-2015. The original GloSea5 forecasts are indicated with dashed lines while RAFT forecasts issued in week 10 are indicated by solid lines.

4 Conclusions and discussion

In a study of long-range forecast skill for weekly summer surface temperatures in Europe, we assess the skill of the UK Met Office's seasonal prediction system GloSea5 against the ERA5 reanalysis. GloSea5 uses a lagged initialization approach where, for the 1993-2015 hindcasts analyzed here, seven members are initialized on the 1st, 9th, 17th and 25th of every month. Our results indicate that the system might benefit from a step-wise model combination approach, where for the earliest forecast lead times, only more re-

cently available runs are used, while a larger set of runs should be employed for lead times beyond two weeks. Furthermore, the forecast skill is increased for lead time beyond two weeks if climatology is included in the ensemble.

For a lagged ensemble system, additional information in the form of observed forecast errors is available for earlier lead times of the older ensemble members. Using the recently proposed RAFT adjustment approach (Schuhen et al., 2020), we have investigated the use of this information to post-process the older members before the forecast is issued. Our results indicate that the application of the RAFT adjustment can improve the RMSE skill of the forecast by as much as 10% compared to climatology. In each time step, the length of the RAFT adjustment period depends on the number of future lead times where the forecast error is expected to correlate with the most recently observed forecast error. We find that the length of the adjustment period varies over time, with a higher correlation across lead times in July and August than in the earlier part of our study period in May and June.

As argued by e.g. Kharin and Zwiers (2003) and Van Schaeybroeck and Vannitsem (2018), the small samples sizes available for seasonal forecasts (23 seasons in our case) require simple post-processing methods in order to avoid overfitting. The RAFT approach is a fairly simple post-processing method whose strength lies in the use of new, otherwise unused, information. The current study focuses on average skill in predicting mean weekly summer temperatures in Europe. For many forecast users, a particularly valuable information is the occurrence of outliers, e.g. a particularly warm or cold summer. While this topic requires further investigation, we expect that RAFT could prove particularly useful in such situations when the outlier has been detected in the newest runs with that not being the case for the older runs.

Acknowledgments

The authors acknowledge the support of the Research Council of Norway: N. Schuhen through grant nr. 259864 “Stipendiatstillinger til Norsk Regnesentral”, T. L. Thorarinsdottir and A. Lenkoski through grant nr. 270733 “Seasonal Forecasting Engine”.

References

Copernicus Climate Change Service (C3S) (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Copernicus Climate Change Service Climate Data Store. https://cds.climate.copernicus.eu/cdsapp#!/home, accessed in November 2019. 5

Doblas-Reyes, F., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L., and Van Oldenborgh, G. (2013a). Initialized near-term regional climate change prediction. *Nature Communications*, 4:1715. 4

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R.

(2013b). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4(4):245–268. 4

Heinrich, C., Hellton, K. H., Lenkoski, A., and Thorarinsdottir, T. L. (2019). Multivariate postprocessing methods for high-dimensional seasonal weather forecasts. arXiv:1907.09716. 4

Hoskins, B. (2013). The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139(672):573–584. 4

Kharin, V. V. and Zwiers, F. W. (2003). Improved seasonal probability forecasts. *Journal of Climate*, 16(11):1684–1701. 11

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., Gordon, M., Vellinga, M., Williams, A., Comer, R., Camp, J., Xavier, P., and Madec, G. (2015). Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, 141(689):1072–1084. 4, 5

Ogallo, L., Bessemoulin, P., Ceron, J.-P., Mason, S., and Connor, S. J. (2008). Adapting to climate variability and change: the climate outlook forum process. *Bulletin of the World Meteorological Organization*, 57(2):93–102. 4

Robertson, A. and Vitart, F., editors (2018). *Sub-seasonal to Seasonal Prediction: The Gap Between Weather and Climate Forecasting*. Elsevier, Amsterdam, Netherlands. 4

Roulin, E. and Vannitsem, S. (2019). Post-processing of seasonal predictions – case studies using EUROSIP hindcast data base. *Nonlinear Processes in Geophysics*, in review. 4

Royer, J. (1993). Review of recent advances in dynamical extended range forecasting for the extratropics. In Shukla, J., editor, *Prediction of Interannual Climate Variations*, pages 49–69. Springer, Berlin, Heidelberg. 4

Schuhen, N. (2019). Order of operation for multi-stage post-processing of ensemble wind forecast trajectories. *Nonlinear Processes in Geophysics*, accepted. 7

Schuhen, N., Thorarinsdottir, T. L., and Lenkoski, A. (2020). Rapid adjustment and post-processing of temperature forecast trajectories. *Quarterly Journal of the Royal Meteorological Society*, in press. 4, 6, 7, 11

Thorarinsdottir, T. L. and Schuhen, N. (2018). Verification: assessment of calibration and accuracy. In Vannitsem, S., Wilks, D. S., and Messner, J. W., editors, *Statistical postprocessing of ensemble forecasts*, chapter 6, pages 155–186. Elsevier, Amsterdam, Netherlands. 5

Van Schaeybroeck, B. and Vannitsem, S. (2018). Postprocessing of long-range forecasts. In Vannitsem, S., Wilks, D. S., and Messner, J. W., editors, *Statistical Postprocessing of Ensemble Forecasts*, chapter 10, pages 267–290. Elsevier, Amsterdam, Netherlands. 4, 11

Vitart, F., Robertson, A. W., and Anderson, D. L. (2012). Subseasonal to seasonal prediction project: Bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization*, 61(2):23. 4