Solánská, J., Kliegrová, S., 2026. Statistical Postprocessing of Long-Range Air Temperature Forecasts in the Czech Republic Using Neural Networks. In: Tolasz, R., Šustková, V. (Ed.), Proceedings of the 2nd PERUN Project Conference (TA ČR, SS02030040). CHMI, Praha, s. 62–68, ISBN 978-80-7653-095-9. https://doi.org/10.59984/978-80-7653-095-9.08

Statistical Post-Processing of Long-Range Air Temperature Forecasts in the Czech Republic Using Neural Networks

(Statistické zpracování dlouhodobých předpovědí teploty vzduchu v České republice s využitím neuronových sítí)

Solánská, J.

https://doi.org/10.59984/978-80-7653-095-9.08

Jana Solánská, Czech Hydrometeorological Institute, jana.solanska@chmi.cz

Stanislava Kliegrová, Czech Hydrometeorological Institute, stanislava.kliegrova@chmi.cz

Abstract: Long-range weather forecasts are an important tool for strategic planning in a number of sectors, particularly in agriculture, energy, and water management. However, in Central Europe – especially in the Czech Republic – their reliability and usability face a number of challenges. This paper focuses on the potential of statistical post-processing of long-range air temperature forecasts in the Czech Republic. We analyzed joint hindcast data from 1993 to 2016 from four global seasonal forecasting systems available through the Copernicus Climate Change Service (Météo-France 8, ECMWF SEAS5, DWD GCFS2.1, and CMCC SPS3.5). Multilayer perceptron neural networks were used for post-processing, with monthly average air temperatures and statistical characteristics of sea-level pressure from the model outputs serving as inputs. The objective was to forecast the tertile categories of monthly average temperature (below normal, near normal, above normal). The Brier Skill Score (BSS) relative to a climatological reference was used to evaluate the quality of the forecasts (both the direct outputs of global forecast models and the results obtained through post-processing using neural networks). The results show that neural networks improve forecasting performance, particularly for below-normal and above-normal temperature categories, with the most significant improvement observed during the winter season. In contrast, the prediction of near-normal conditions remains limited. The study confirms the potential of machine learning methods to enhance the usability of long-range forecasts in the Czech Republic.

Keywords: Long-range weather forecast – neural networks – statistical post-processing

Abstrakt: Dlouhodobé předpovědi počasí jsou důležitým nástrojem pro strategické plánování v řadě sektorů, zejména v zemědělství, energetice a vodním hospodářství. V podmínkách střední Evropy, zejména v České republice, však jejich spolehlivost a použitelnost čelí řadě výzev. Tento příspěvek se zaměřuje na potenciál statistického post-processingu dlouhodobých předpovědí teploty vzduchu v České republice. Analyzována byla společná hindcastová data z období 1993–2016 ze čtyř globálních sezónních předpovědních systémů dostupných v rámci Copernicus Climate Change Service (Météo-France 8, ECMWF SEAS5, DWD GCFS2.1 a CMCC SPS3.5). K post-processingu byly využity vícevrstvé perceptronové neuronové sítě, jejichž vstupy tvořily měsíční průměry teploty vzduchu a statistické charakteristiky tlaku na hladině moře z modelových výstupů. Cílem bylo předpovědět tercilové kategorie měsíční průměrné teploty (pod normálem, blízko normálu, nad normálem). K hodnocení kvality předpovědí (jak přímých výstupů globálních předpovědních modelů, tak výsledků získaných post-processingem pomocí neuronových sítí) bylo použito Brier Skill Score (BSS) vzhledem ke klimatologické referenci. Výsledky ukazují, že neuronové sítě zvyšují předpovědní dovednost zejména pro kategorie podnormálních a nadnormálních teplot, přičemž nejvýraznější zlepšení bylo

zaznamenáno v zimním období. Naopak předpověď podmínek blízkých normálu zůstává omezená. Studie potvrzuje potenciál metod strojového učení pro zvýšení využitelnosti sezónních předpovědí v podmínkách České republiky.

Klíčová slova: dlouhodobá předpověď počasí – neuronové sítě – statistický post-processing

Introduction

Long-range forecasts serve as a bridge between conventional weather forecasts and climate projections. While standard weather forecasts focus on specific weather conditions in the coming days, long-range forecasts assess the likely evolution of meteorological conditions over a period of weeks to months. These forecasts are important for agriculture (Bruno Soares 2017), water management (Svensson et al. 2015), energy (Clark et al. 2017; De Felice et al. 2015), transportation (Palin et al. 2016), and emergency management. They provide valuable information about the likely nature of the coming period.

In long-range forecasts, components of the climate system that change more slowly than the atmosphere itself play a significant role. These include ocean temperatures, snow cover, soil moisture, sea ice, and large-scale atmospheric circulation. It is precisely these factors that can influence the weather over a period of several weeks or months and provide a source of predictability. For example, the El Niño phenomenon in the Pacific Ocean can influence the weather in many parts of the world for several months. Long-range forecasts for Central Europe remain very challenging. This is due to the high variability of atmospheric processes, the region's complex geographical location, and relatively weak climate signals outside the winter season.

As part of the PERUN project, we investigated the use of neural networks for the post-processing of long-range air temperature forecasts for the Czech Republic. The main objective was to determine how different neural network architectures and data resampling strategies affect the resulting forecast quality.

Data

Data from four major European modeling centers were used for the analysis: European Centre for Medium-Range Weather Forecasts (ECMWF), Meteo-France, Deutscher Wetterdienst (DWD), and the Italian Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC). Hindcasts covering the period 1993–2016 were available. The study area included the territory of the Czech Republic, represented by 24 grid points with a spatial resolution of 1° (49–51° N, 12–19° E). The homogenized GriSt (Beranová 2024) temperature series was used as reference data and served as the basis for verifying the quality of the model forecasts. Monthly temperature values, categorized into three groups using tertiles of the climatological distribution – below normal, normal, and above normal – were used as the predictand.

Method

Multilayer perceptron (MLP) neural networks were used for statistical analysis. The input layer consisted of 21 neurons – the monthly average of air temperature and sea-level pressure from four forecast models, the variability of air temperature and sea-level pressure represented by the difference between the 10th and 90th percentiles from four forecast models, latitude and longitude, forecast lead time, and forecast month, expressed as continuous functions using sine and cosine.

Forecasts were generated for lead times of one to three months. Two data resampling strategies were tested: bootstrap and random data redistribution. Bootstrap resampling generated alternative realizations of the training dataset through sampling with replacement. Each bootstrap sample retained the same size as the original dataset and was subsequently divided into training, validation, and testing subsets using a fixed ratio of 70% / 15% / 15%. This approach increases ensemble diversity and allows estimation of predictive uncertainty arising from sampling variability.

The random resampling approach consisted of repeated random partitioning of the original dataset into training, validation, and testing subsets using the same fixed ratio but without replacement. For each neural network within the ensemble, an independent random split was generated, allowing individual

models to be trained on slightly different subsets of the available data. Compared with bootstrap resampling, this strategy preserves the original structure of the dataset and avoids duplicated samples in the testing subset.

For both resampling approaches, ensembles of 100 neural networks were generated.

At the same time, two approaches to model outputs were compared – processing individual categories separately and processing all three categories simultaneously. The accuracy of the forecasts was evaluated using the Brier Skill Score (BSS), which measures improvement over the reference forecast, where positive BSS values indicate better model performance.

The architecture of the multi-output neural network is shown in Fig. 1.

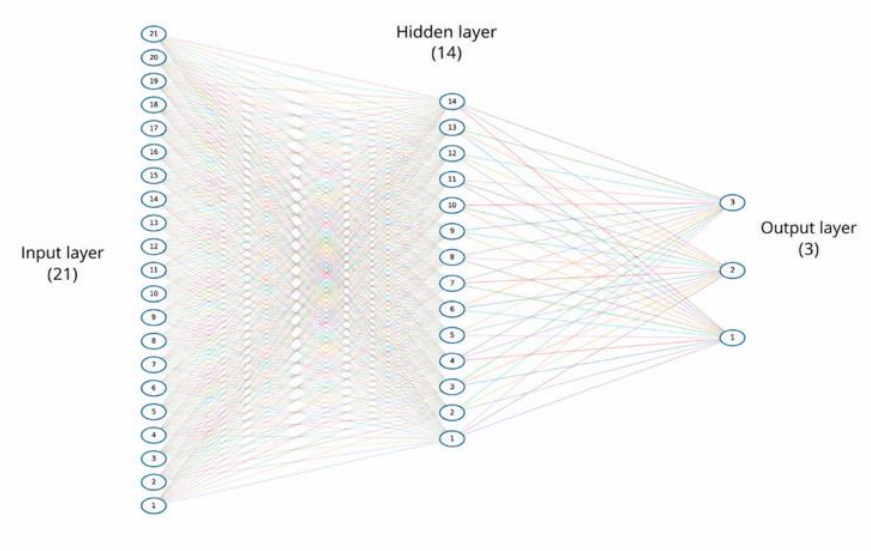


Fig. 1. Structure of a Neural Network with multiple outputs.

Discussion

We compared two variants of neural networks. The first approach used separate binary classifiers for each temperature category. Each network, therefore, predicted only the probability of a single specific category. The second approach used a single neural network with multiple outputs, which simultaneously predicted the probabilities of all three categories. Thanks to the use of the Softmax function, the resulting probabilities were consistent with one another, and their sum was always equal to one.

The results showed that the multi-output network achieved better results than the separate binary classifiers (Fig. 2). While models focused on individual categories often degraded the quality of predictions compared to the original model outputs, the multi-output network generally improved the predictions or at least maintained their quality. The greatest improvement was observed in the below-normal and above-normal temperature categories. The results suggest that when predicting tertile categories, it is important to preserve their interrelationships and predict them jointly. Dividing the problem into several independent binary tasks leads to a loss of information and a reduction in the quality of the results.

Another aim of the study was to compare two resampling strategies used in training neural networks. The first method was bootstrapping, in which data are selected with replacement. This approach creates greater variability in the training sets and can increase the model's resistance to overfitting. The second method was random resampling without replacement. During each training run, the data was redistributed into training, validation, and test sets, but without creating duplicate records.

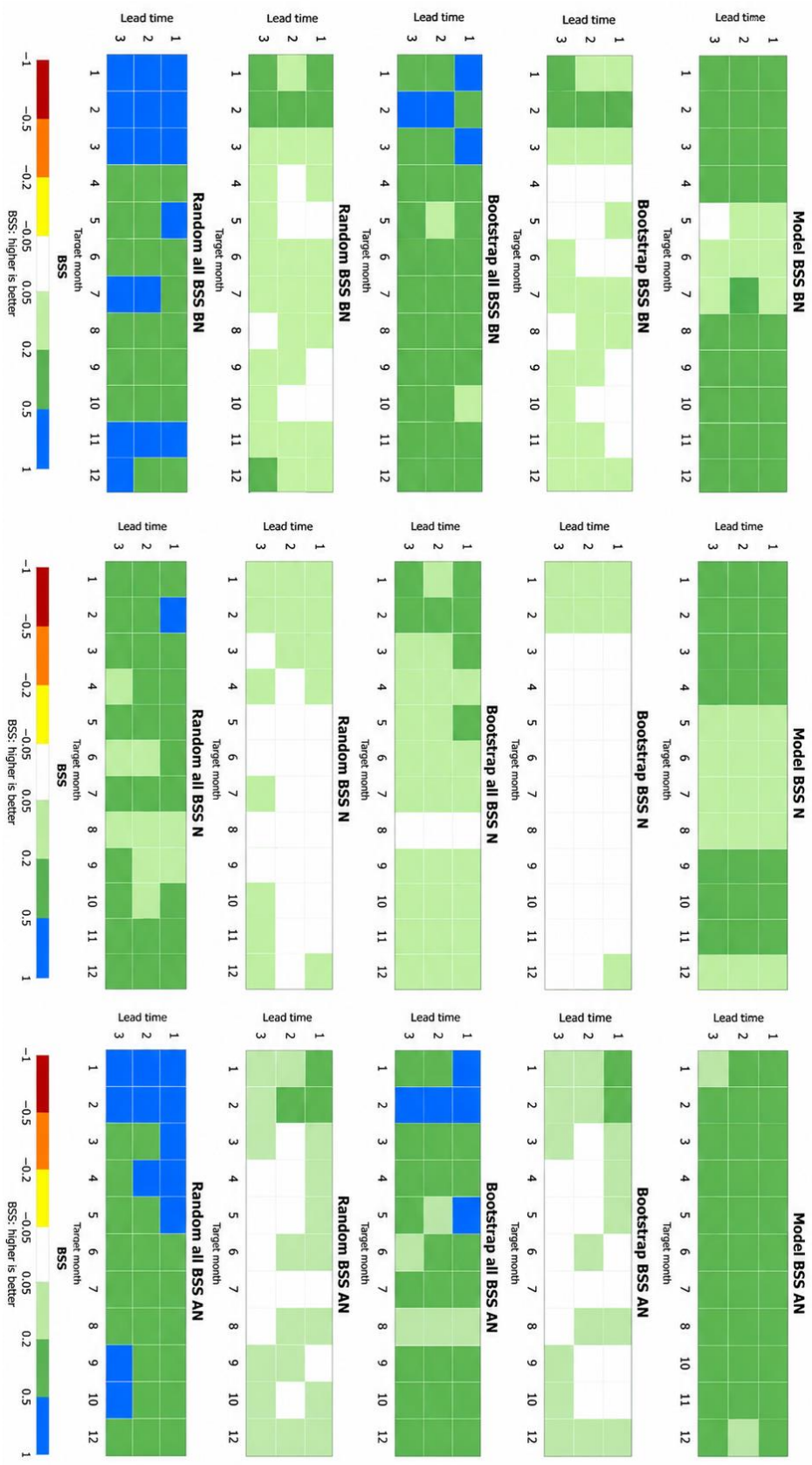


Fig. 2. Heat maps of the Brier Skill Score (BSS) for forecast months and lead times. The top row shows BSS values for the raw model forecast, rows show BSS values for the neural network post-processed forecast for below-normal categories (BN), near-normal categories (N) and above-normal categories (AN).

The results showed that the influence of the resampling strategy was smaller than that of the neural network architecture itself. Nevertheless, random resampling was found to generally provide more stable and spatially consistent results (Fig. 3). Although bootstrap resampling sometimes achieved very good results, particularly during the winter, it exhibited greater variability and less stable behaviour across individual months and forecast horizons. For practical use in operational settings, random resampling therefore appears to be more suitable.

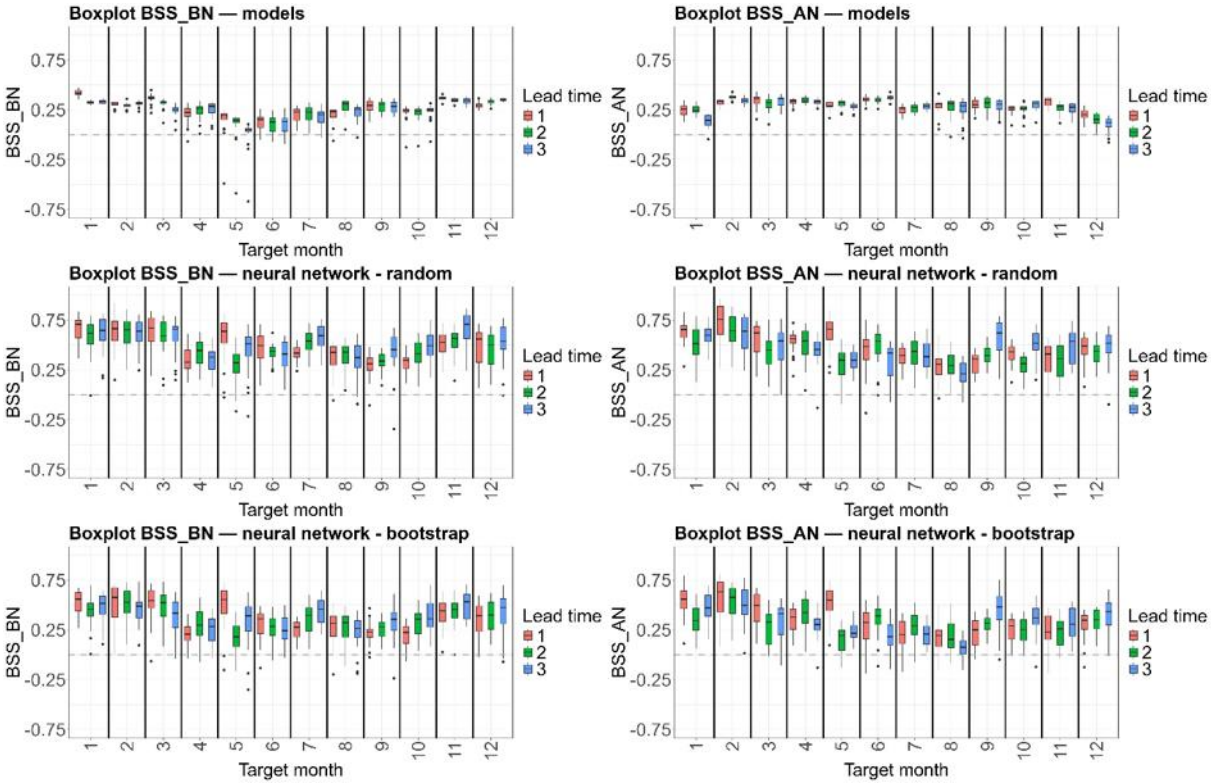


Fig. 3. Boxplots of the Brier Skill Score (BSS) for the raw model forecast and for the neural network post-processed forecast with bootstrap resampling or random resampling month–lead time combinations, shown separately for below-normal (BN) and above-normal (AN) categories for the raw model forecasts. The box shows the interquartile range (25^{th}–75^{th} percentiles) with the median marked by a horizontal line. Whiskers extend to 1.5 times the interquartile range, and points beyond this range indicate outliers.

A very significant finding of the study was the strong dependence of forecast accuracy on the season. The highest Brier Skill Score (BSS) values, which indicate the quality of probabilistic forecasts, were achieved during the winter months and in early spring. This result is related to physical processes in the atmosphere. During the winter, the European climate is influenced by strong, large-scale circulation systems, such as the North Atlantic Oscillation (NAO) or interactions between the stratosphere and the troposphere (Hurrell et al. 2003; Scaife et al. 2014, Domeisen et al. 2018). These phenomena provide forecast models with a relatively strong and stable signal, which neural networks can further utilize and refine. In contrast, weather predictability is significantly lower in the summer. Atmospheric processes are more influenced by local factors, and the signal-to-noise ratio decreases. Under these conditions, neural networks are no longer able to significantly improve results and can sometimes even lead to worse forecasts. Forecasting the normal temperature category proved to be the most difficult. This category exhibited the lowest BSS values across all configurations, confirming that weak temperature deviations close to the long-term normal are difficult to predict (Barnston and Tippett 2014; Calì Quaglia et al. 2022).

The study also showed that neural networks influence the spatial structure of forecasts. While the original model outputs were relatively homogeneous, the post-processed results exhibited greater spatial

variability. An interesting finding was the deterioration in accuracy in the peripheral parts of the study area (Fig. 4). This effect is likely due to the fact that points at the domain boundaries have less spatial information available from the surrounding area than points in the center of the region. In contrast, higher forecast accuracy values were often achieved in the central part of the Czech Republic. This finding suggests that in future method development, it will be advisable to work with a broader geographic area to limit negative edge effects.

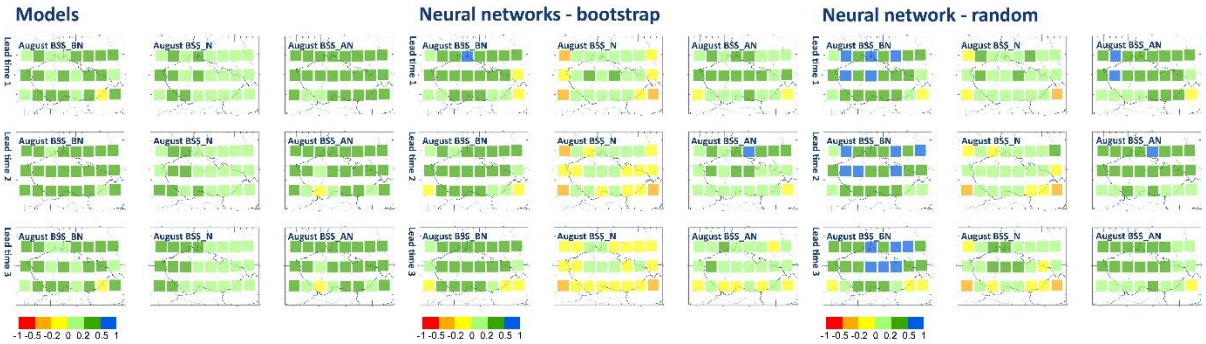


Fig. 4. Brier Skill Score (BSS) values of the long-range probabilistic forecast for August, shown separately for the below-normal (BSS_BN), near-normal (BSS_N), and above-normal (BSS_AN) categories. Results are displayed for three lead times, corresponding to forecasts initialized on 1 May (lead time 3), 1 June (lead time 2), and 1 July (lead time 1).

Conclusion

The study demonstrated that neural networks can be a useful tool for improving long-range air temperature forecasts in the Czech Republic. The best results are achieved by multi-output neural networks, which simultaneously forecast all temperature categories and preserve their interrelationships.

It also turned out that the choice of neural network architecture is more important than the choice of resampling strategy. Random resampling provides slightly more stable results than bootstrap, but the differences between the two methods are not significant.

The results further confirm that the success of machine learning methods is closely linked to the physical predictability of the atmospheric system. Neural networks are capable of utilizing existing forecast signals and improving their interpretation; however, they cannot, on their own, create new predictability where it is lacking in the data.

These results demonstrate the potential of artificial intelligence for operational long-range forecasting and provide a basis for further methodological development.

Acknowledgments

The research was partially funded by the Technology Agency of the Czech Republic (grant no. SS02030040, PERUN).

References

Barnston, A. G., Tippett, M. K., 2014. Climate information, outlooks, and understanding – where does IRI stand? *Earth Perspectives*, **1**(20). https://doi.org/10.1186/2194-6434-1-20

Beranová, R., et al., 2024. Comparison of PERUN/Reanalysis with station data in the regular network. *Meteorologické zprávy*, **77**(3). https://doi.org/10.59984/mz.2024.03.01

Bruno Soares, M., 2017. Assessing the usability and potential value of seasonal climate forecasts in land management decisions in the southwest UK. *Wea. Climate Soc.*, **9**, pp. 167–181. https://doi.org/10.5194/asr-14-175-2017

Calì Quaglia, F., et al., 2022. Temperature and precipitation seasonal forecasts over the Mediterranean region: Added value compared to simple forecasting methods. *Climate Dynamics*, **58**, pp. 2167–2191. https://doi.org/10.1007/s00382-021-05895-6

Clark, R. T., et al., 2017. Skilful seasonal predictions for the European energy industry. *Environ. Res. Lett.*, **12**, 024002. https://doi.org/10.1088/1748-9326/aa57ab

De Felice, M., et al., 2015. Seasonal climate forecasts for medium-term electricity demand forecasting. *Appl. Energy*, **137**, pp. 435–444. https://doi.org/10.1016/j.apenergy.2014.10.030

Domeisen, D. I. V., et al., 2018. How predictable are the Arctic and North Atlantic Oscillations? *J. Climate*, **31**, pp. 997–1014. https://doi.org/10.1175/JCLI-D-17-0226.1

Hurrell, J. W., et al., 2003. An overview of the North Atlantic Oscillation. *Geophys. Monogr.*, **134**, Amer. Geophys. Union, pp. 1–35. https://doi.org/10.1029/134GM01

Palin, E. J., et al., 2016. Skillful seasonal forecasts of winter disruption to the U. K. transport system. *Journal of Applied Meteorology and Climatology*, **55**, pp. 325–344. https://doi.org/10.1175/JAMC-D-15-0102.1

Scaife, A. A., et al., 2014. Skillful long-range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, pp. 2514–2519, https://doi.org/10.1002/2014GL059637

Svensson, C, et al., 2015. Long-range forecasts of UK winter hydrology. *Environmental Research Letters*, **10**, 064006. https://doi.org/10.1088/1748-9326/10/6/064006