BAMS In Box

Outcomes of the WMO Prize Challenge to Improve Subseasonal to Seasonal Predictions Using Artificial Intelligence

F. Vitart, A. W. Robertson, A. Spring, F. Pinault, R. Roškar, W. Cao, S. Bech, A. Bienkowski, N. Caltabiano, E. De Coning, B. Denis, A. Dirkson, J. Dramsch, P. Dueben, J. Gierschendorf, H. S. Kim, K. Nowak, D. Landry, L. Lledó, L. Palma, S. Rasp, and S. Zhou

ABSTRACT: There is a high demand and expectation for subseasonal to seasonal (S2S) prediction, which provides forecasts beyond 2 weeks, but less than 3 months ahead. To assess the potential benefit of artificial intelligence (AI) methods for S2S prediction through better postprocessing of ensemble prediction system outputs, the World Meteorological Organization (WMO) coordinated a prize challenge in 2021 to improve subseasonal prediction. The goal of this competition was to produce the most skillful forecasts of precipitation and 2-m temperature globally averaged over forecast weeks 3 and 4 and over weeks 5 and 6 for the year 2020 using artificial intelligence techniques. The top three submissions, described in this article, succeeded in producing S2S forecasts significantly more skillful than the bias-corrected ECMWF operational reference forecasts, particularly for precipitation, through improved calibration of the ECMWF raw forecast outputs or multimodel combination. These forecast improvements should benefit the use of S2S forecasts in applications.

KEYWORDS: Neural networks; Regression analysis; Statistical techniques; Forecast verification/skill; Numerical weather prediction/forecasting; Model evaluation/performance

https://doi.org/10.1175/BAMS-D-22-0046.1

Corresponding author: Frédéric Vitart, frederic.vitart@ecmwf.int

In final form 22 September 2022

AFFILIATIONS: Vitart, Pinault, Dramsch, and Dueben—European Centre for Medium-Range Weather Forecasts, Reading, Berkshire, United Kingdom; **Robertson**—International Research Institute for Climate and Society, Columbia University, Palisades, New York; **Spring**—Max Planck Institute for Meteorology, Hamburg, Germany; **Roškar**—Swiss Data Science Center, Zurich, Switzerland; **Cao, Caltabiano, and De Coning**—World Meteorological Organization, Geneva, Switzerland; **Bech, Lledó, and Palma**—Barcelona Supercomputing Center, Barcelona, Spain; **Bienkowski, Kim, and Zhou**—University of Connecticut, Storrs, Connecticut; **Denis**—Montreal, Quebec, Canada; **Dirkson**—Environment and Climate Change Canada, Dorval, Quebec, Canada; **Gierschendorf and Landry**—Computer Research Institute of Montreal, Montreal, Quebec, Canada; **Nowak**—Boulder Canyon Operations, Lower Colorado Region, Office of Reclamation, Boulder City, Nevada; **Rasp**—ClimateAI, Inc., San Francisco, California

Skillful subseasonal to seasonal (S2S) prediction is important to inform decision-makers regarding, for example, changes in risks of extreme events or opportunities for optimizing resource management. However, the skill at this time range is often low, or only marginally better than climatology or persistence forecasts (Robertson et al. 2020), and depends strongly on the presence of active sources of S2S predictability ("windows of opportunity") (Mariotti et al. 2020). To improve forecast skill and understanding on the S2S time scale, the World Weather Research Programme (WWRP) and World Climate Research Programme (WCRP) launched the Subseasonal to Seasonal Prediction Project (S2S) (Vitart et al. 2015) in November 2013.

Artificial intelligence (AI) can potentially improve S2S predictions because of its potential to explore very large multimodel forecast and observed datasets more agnostically, to discover emergent patterns in the data (e.g., Weyn et al. 2021), instead of first reducing them a priori to limited subspaces and variables as is traditionally done. However, a major challenge is the small size of model reforecast data used to train the AI method. Therefore, AI was identified as a key research topic in weather and climate science for the upcoming years by WMO. A prize challenge to improve S2S predictions using AI was organized by the WWRP–WCRP S2S project in collaboration with the Swiss Data Science Center (SDSC) and the European Centre for Medium-Range Weather Forecasts (ECMWF). This competition fostered this approach by specifically encouraging the use of AI tools to extract valuable information from the large S2S database developed by the S2S project (Vitart et al. 2017). This database contains real-time forecasts and reforecasts from 12 operational S2S models and provides a unique opportunity for assessing the benefits of AI methods for subseasonal prediction. AI can be based on various methods with versatile options for model input, architectures, and loss functions. The WMO S2S AI Challenge provided a structured means to explore different options in a comparable way across the S2S and AI/ML communities.

The S2S Artificial Intelligence Challenge

The goal of the S2S AI Challenge was to provide the "best possible" probabilistic forecast of temperature and precipitation for weeks 3 and 4 (days 15–28) and weeks 5 and 6 (days 29–42), henceforth week 3 + 4 and week 5 + 6. The forecast domain was global, on a 1.5° spatial grid resolution but limited to land grid points. The forecasts were requested to be issued each Thursday of the year 2020 in the form of tercile-category probabilities, following

standard forecasting practice. The participants were required to develop, train, and test their AI models using only observed data and reforecasts (e.g., 20 years and 11 ensemble members for the ECMWF reforecasts) prior to 2020, and to submit their forecasts for the year 2020. The verification was performed using the ranked probability skill score (RPSS) (Epstein 1969) on four domains: global, northern extratropics, tropics, and southern extratropics using observed gridded data from NOAA Climate Prediction Center (CPC): the CPC unified gauge-based analysis of global daily precipitation (Chen et al. 2008) and the CPC Global Unified Temperature (available from http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP/.CPC/.temperature/.daily/). This competition was similar to the S2S Rodeo (www.usbr.gov/research/challenges/forecastrodeo.html), but not limited to a specific region. The created software, code, documentation, and results were required to be open source and open access. More details on the rules of the competition can be found at https://s2s-ai-challenge.github.io/. The competition encouraged the use of AI methods on forecasts and reforecasts from models in the S2S database to provide better calibration and multimodel ensemble combination, but AI methods designed to replace dynamical models were also welcome. There was no constraint on the AI methodology.

The competition took place between 1 June and 1 November 2021. The top three winners, from Computer Research Institute of Montreal (CRIM), Barcelona Supercomputing Center (BSC), and University of Connecticut (UConn), were announced at a webinar in February 2022 (recording available at www.s2sprediction.net/static/webinar). A prize of 30,000 Swiss francs was shared among them.

Technical setup

An important aspect of this challenge was to provide a convenient technical environment for the competitors, allowing most of their time to be spent on developing and testing methods, rather than on manipulating data. This was greatly facilitated by the "Renku" platform from the SDSC and the European Weather Cloud from ECMWF.

Bootstrapping the competition on the Renku platform. "Renku" is a platform for enabling reproducible and collaborative data science projects developed by the Swiss Data Science Center at the ETH Zürich and EPFL. For the S2S competition, the public platform at https://renkulab.io was used to host the code repositories and provide a simple framework for participants to get started quickly with the competition, including data ingestion, basic model building, and an automatic scoring system. Participants could immediately launch interactive sessions on the Renku platform to set up their projects and experiment with sample data.

The EWC infrastructure and CliMetLab. The European Weather Cloud (EWC) is a cloud infrastructure established by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) and ECMWF. It provided fast and easy access to a subset of the S2S database under multiple formats for the competition. This included daily and biweekly real-time forecasts and reforecasts from three S2S models: ECMWF, Environment and Climate Change Canada (ECCC), and the National Centers for Environmental Prediction (NCEP). In addition, access to virtual machines on the EWC was offered to participants from developing countries. The S2S data access from the cloud was facilitated by the use of an open-source Python software called CliMetLab (https://climetlab.readthedocs.io).

Organization of the competition. Four invited experts reviewed the submissions for compliance with the competition rules. At the end of the peer review period, all the submissions beating climatology were presented in a virtual session.

Results from the competition

Of the 47 teams who registered for the competition, 10 provided a submission. Five of these provided a global RPSS score higher than climatological odds and ECMWF forecasts, which had been bias corrected by computing the tercile thresholds from the model's reforecast climatology rather than from observations. The top three submissions provided skillful 2020 forecasts of temperature and precipitation (Figs. 1a,b, red color) over larger portions of the globe than the ECMWF reference forecasts, although there are areas where the latter are more skillful (e.g., 2-m temperature over northern Asia). The top entry (CRIMS2S) produced substantially more skillful precipitation forecasts than the ECMWF reference (Fig. 1b), with fewer areas of negative skill. Table 1 shows that the top three entries produced better forecasts than climatology and ECMWF forecasts over most areas, with an average global RPSS of 0.046 for CRIMS2S, 0.029 for BSC, and 0.006 for UConn compared to the ECMWF benchmark (−0.001) and climatology (0).

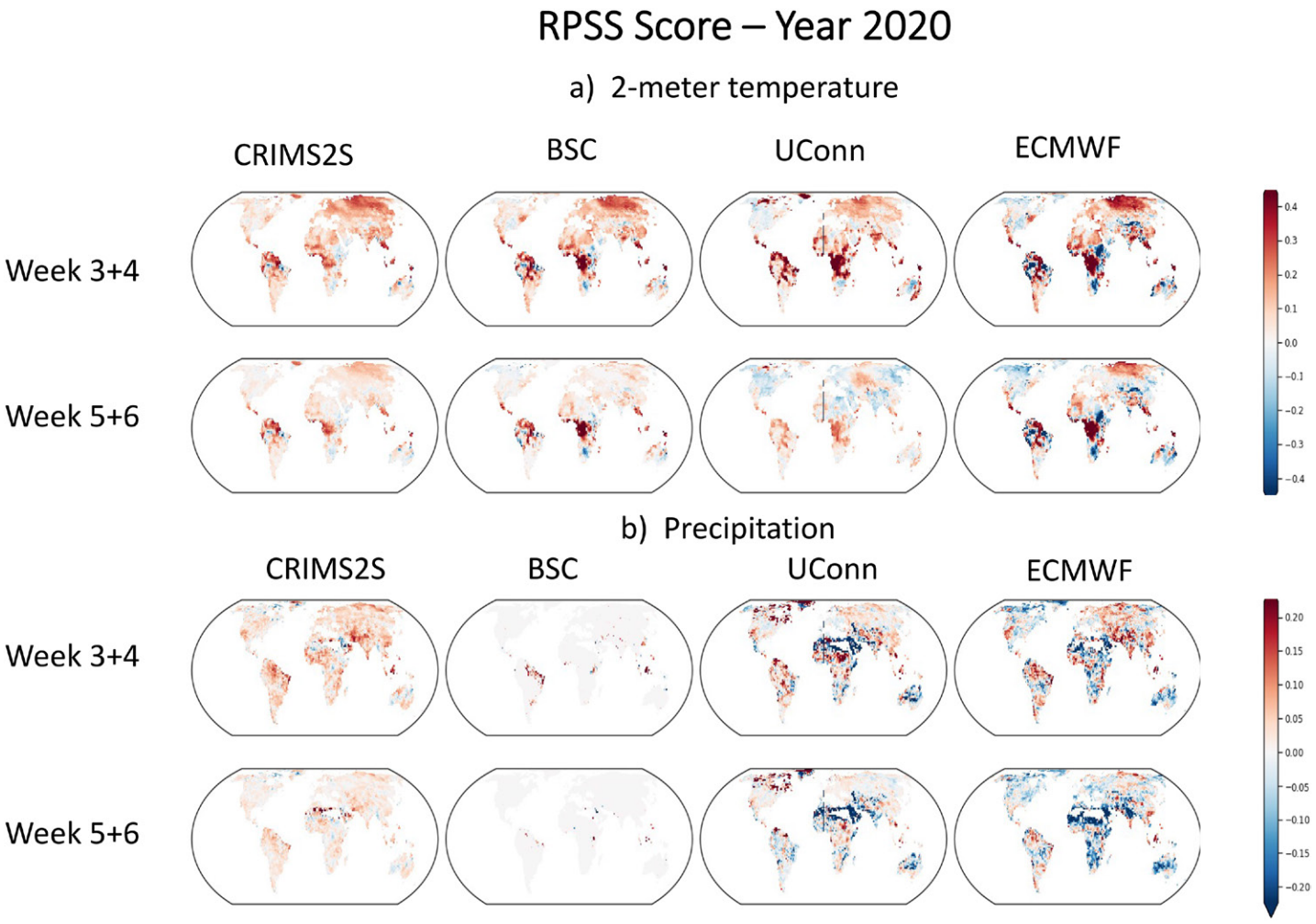

Fig. 1. Ranked probabilistic skill score (RPSS) maps of (a) 2-m temperature and (b) precipitation for week 3 + 4 (top row in each panel) and week 5 + 6 (bottom row in each panel). The columns represent the performances of (first column) CRIMS2S, (second column) BSC, and (third column) UConn. (fourth column) The performance of the ECMWF model with simple bias correction, as a reference. Red (blue) colors indicate better (worse) skill than climatology. The RPSS compares the ranked probabilistic score (RPS) of the forecast to the RPS of climatology, so that negative (positive) values indicate worse (better) performances than climatology. Better performances are indicated by higher values of the RPSS.

Table 1. RPSS of 2-m temperature and precipitation computed over several regions and biweekly time ranges for the top three entries.

		2-m temperature				Total precipitation			
Lead time	Model	90°–30°N	30°N–30°S	30°S–90°S	Global	90°–30°N	30°N–30°S	30°S–90°S	Global
Week 3 + 4	CRIMS2S	0.099	0.101	0.037	0.090	0.027	0.042	0.029	0.030
	BSC	0.080	0.107	0.044	0.082	0.000	0.006	0.004	0.003
	UConn	0.063	0.129	0.074	0.080	0.010	−0.062	0.008	−0.017
	ECMWF bench	0.053	0.074	0.003	0.045	−0.013	−0.002	−0.014	−0.012
Week 5 + 6	CRIMS2S	0.044	0.064	0.011	0.046	0.013	0.026	0.016	0.017
	BSC	0.017	0.071	−0.002	0.030	−0.000	0.001	−0.000	0.000
	UConn	−0.034	0.025	0.030	−0.013	0.005	−0.088	0.005	−0.030
	ECMWF bench	−0.003	0.045	−0.043	0.000	−0.030	−0.061	−0.012	−0.044

CRIMS2S team[1] contribution to the S2S-AI Challenge. The CRIMS2S team proposed an "opportunistic mixture model." It consisted of a weighted multimodel ensemble based on five predictions: (i) ECMWF, ECCC, and NCEP forecasts, each post-processed using ensemble model output statistics (EMOS) following Gneiting et al. (2005), which is a variant of multiple linear regression; (ii) a prediction based on a convolutional neural network (CNN) applied to ECMWF forecasts; and (iii) climatology. Years 2000–18 were used for training and 2019 for validation. The test set was year 2020 as requested by the challenge dataset. To reduce the negative impacts of overfitting, weekly EMOS parameters are smoothed using a rolling 20-week window centered on the target forecast week. The weights of the mixture model were obtained by an additional CNN that uses ECMWF forecasts as input, and outputs weights for each of the five models. This design is intended to capture the overall weather state determined by the ECMWF forecasts, and then infer the relationship between the current conditions and the optimal weight for each model in the ensemble. Both the postprocessing and the weighting convolutional neural networks have architectures that are largely inspired by ResNet (He et al. 2016), in that they are built with residual blocks containing skip connections.

[1] David Landry, Jordan Gierschendorf, Alan Dirkson, and Bertrand Denis.

Figure 2 shows the relative weight assigned to the climatology forecast, which, as expected, increases with lead time and is lower in the tropical regions. While this mixture model was successful in the challenge, subsequent experiments have indicated that a simple mean of the five input models performs on a par with the mixture model for 2020. Future work is needed to establish the pertinence of larger convolutional models for this task. The convolutional network did not improve the ECMWF forecasts as much as EMOS. However, keeping both models in the multimodel combination was useful since their errors were not fully correlated.

BSC team[2] contribution to the WMO S2S-AI Challenge. The BSC approach was a point-by-point statistical correction of ECMWF forecasts that transforms raw ensemble predictions into calibrated tercile-category probabilities. Four competing methods were trained for each grid point, lead time, and variable. For each variable, the ECMWF ensemble predictions of that variable were used as the sole predictors. First, two classical methods in climate prediction were used: a climatological forecast and a simple counting of the members exceeding the 33rd and 66th percentiles of the reforecasts. Then two machine learning techniques were used: logistic regression and random forest classification (James et al. 2013). Logistic regression is a simple statistical method that

[2] Llorenç Lledó, Sergi Bech, Lluís Palma, Andrea Manrique-Suñén, and Carlos Gómez Gonzalez.

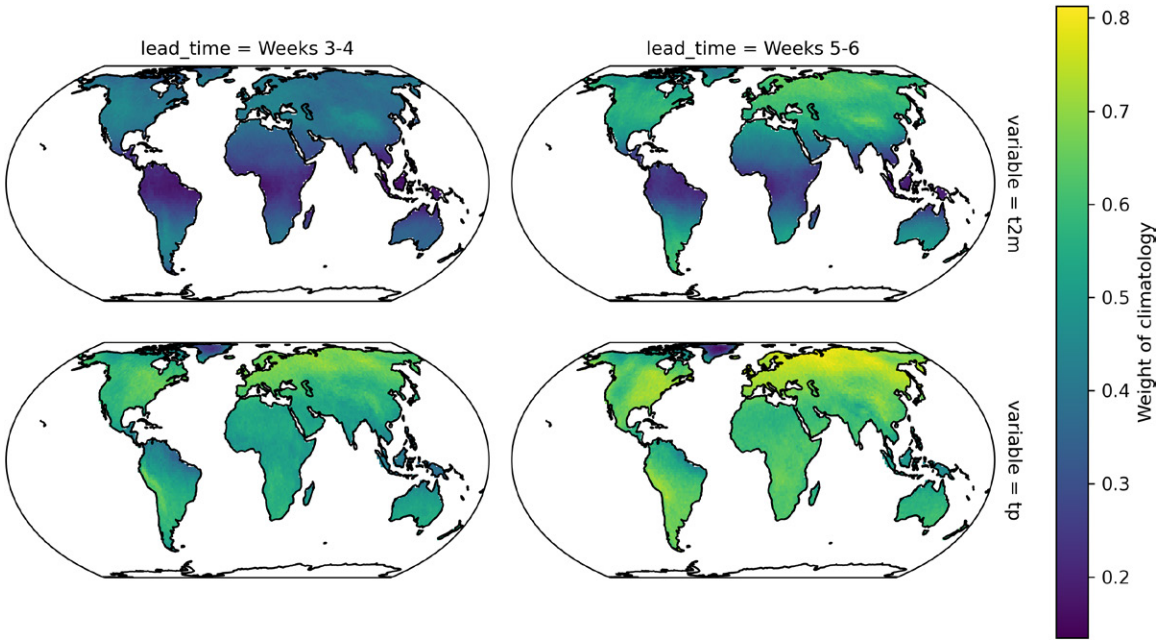

Fig. 2. Mean of the relative weight assigned to climatology by the mixing CNN for (top) surface temperature and (bottom) precipitation forecasts over the test set for (left) week 3 + 4 and (right) week 5 + 6. A lower value indicates more reliance on the dynamical predictions to produce forecasts.

predicts the probabilities of a binary outcome based on observations of a numerical quantity, while random forest classification is a machine learning algorithm that combines the output of multiple decision trees to derive the probabilities of each outcome. For the logistic regression, a one-versus-rest multiclass implementation that models the probabilities of observing each tercile category as a regression on the ECMWF ensemble mean (as in Hamill et al. 2004) was used. To increase the training sample, all weeks of the year were pooled using ensemble-mean anomalies with respect to the biweekly varying climatological mean. The random forest classification used previously sorted ensemble members as features. The method combined the class frequencies observed during training at each leaf by 100 different decision trees of depth 4 to produce the final class probabilities. The quality of the four methods was evaluated in a leave-one-year-out cross validation during the 2000–19 period, and the best method was selected at each grid point based on the median RPSS of all years. Due to time constraints, the two machine learning methods were applied only on temperature. According to Fig. 3, logistic regression performed best over most grid points, but climatology remains unbeaten in some grid points for the week 5 + 6 forecast.

UConn Team[3] contribution to the WMO S2S-AI Challenge. The UConn team's approach began by dividing the forecast area into 23 regions based on approximate ranges of similar climates provided by a weather expert from the Naval Research Laboratory, Monterey, California, as shown in Fig. 4. For each region, lead time, and quantity being forecasted (temperature and precipitation), a random forest classification model from the Python library scikit-learn (Pedregosa et al. 2012) with 100 trees, maximum depth of 10, and Gini impurity criterion was trained and applied to the test data to predict the tercile class, i.e., whether the test observation is below, at, or above normal. Further testing showed that using a separate temperature forecasting model for each location works best, while a single precipitation prediction model suffices for all locations for the 28-day lead time.

[3] Adam Bienkowski, Shanglin Zhou, and Hee-Seung Kim.

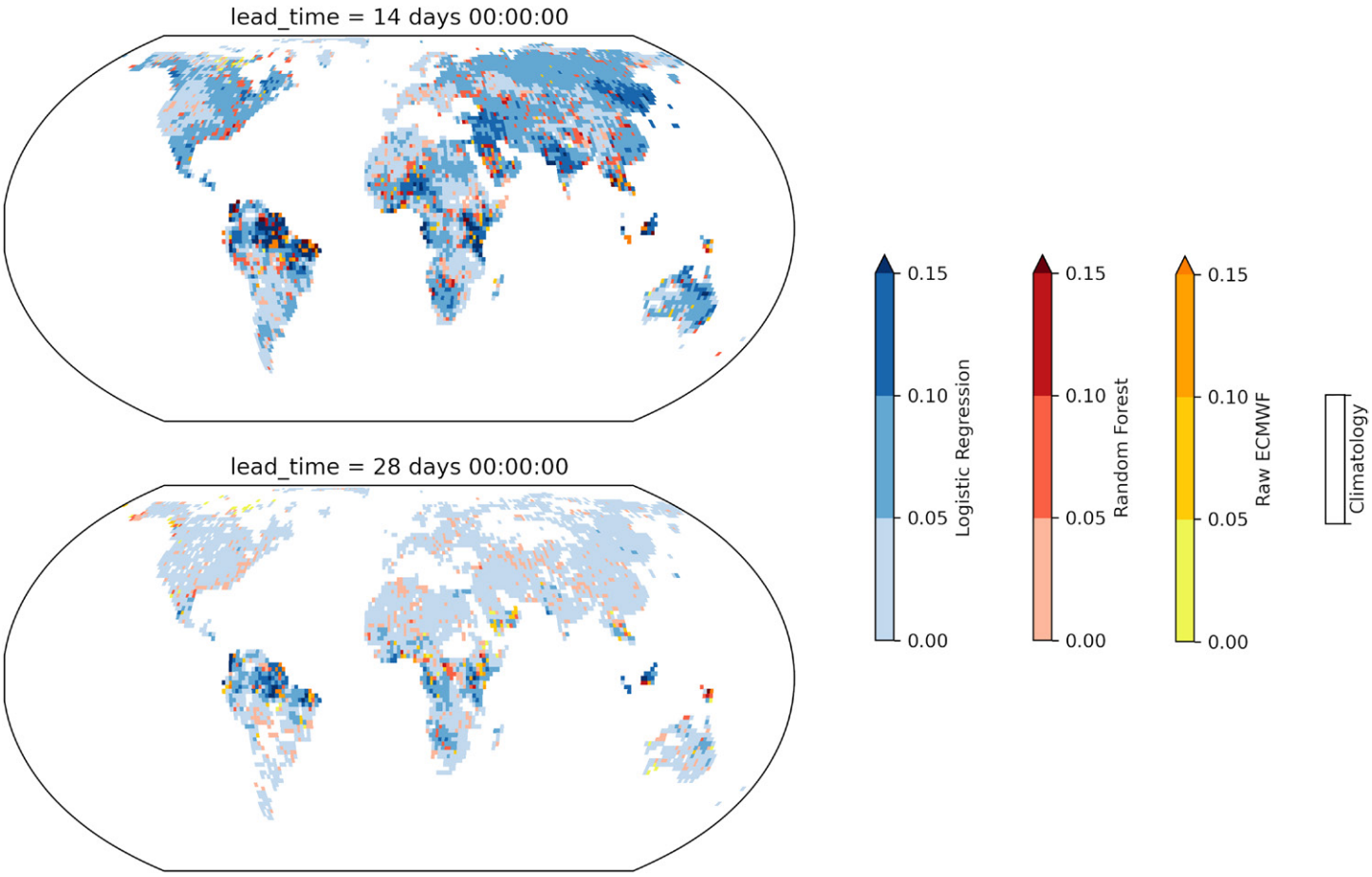

Fig. 3. Median RPSS for 2000–19 temperature forecasts for (top) week 3 + 4 and (bottom) week 5 + 6 for BSC's approach. The colors blue, red, and orange indicates the best model (logistic regression, random forest, and raw ECMWF, respectively) and the intensity shows the skill level.

The period 2000–18 was used for training, 2019 for verification and validation, and 2020 for testing.

The features used in the UConn's models consisted of past observations at the forecast location, in addition to the mean, standard deviation, and median of the observed temperature/precipitation at the forecast location for the same 2-week period over the entire training period, and El Niño–Southern Oscillation (ENSO) indices Niño-1+2 and Niño-3.4 (https://climatedataguide.ucar.edu/climate-data/nino-sst-indices-nino-12-3-34-4-oni-and-tni) for the previous month at the forecast time. The past observations consisted of observed values from the previous 9 days prior to the forecast day, as well as observations from the same day of the year as the forecast day for the past 10 years for week 3 + 4 and past year for week 5 + 6 (this provided better results for week 5 + 6 than using the past 10 years as for week 3 + 4). Further work since the competition showed that using the average of the ECMWF hindcast realizations as a feature in all the models could significantly improve the forecast accuracy by 12.4% for temperature at week 3 + 4 and 5.7% for week 5 + 6.

Conclusions

The AI prize challenge fulfilled its main objective which was to demonstrate that AI methods can be used to provide more skillful S2S forecasts of 2-m temperature and precipitation, compared to simple bias correction. The three winning entries presented a variety of methodologies, including random forest classification (BSC and UConn), convolutional neural networks (CRIMS2S), logistic regression (BSC), and EMOS multiple linear regression (CRIMS2S).

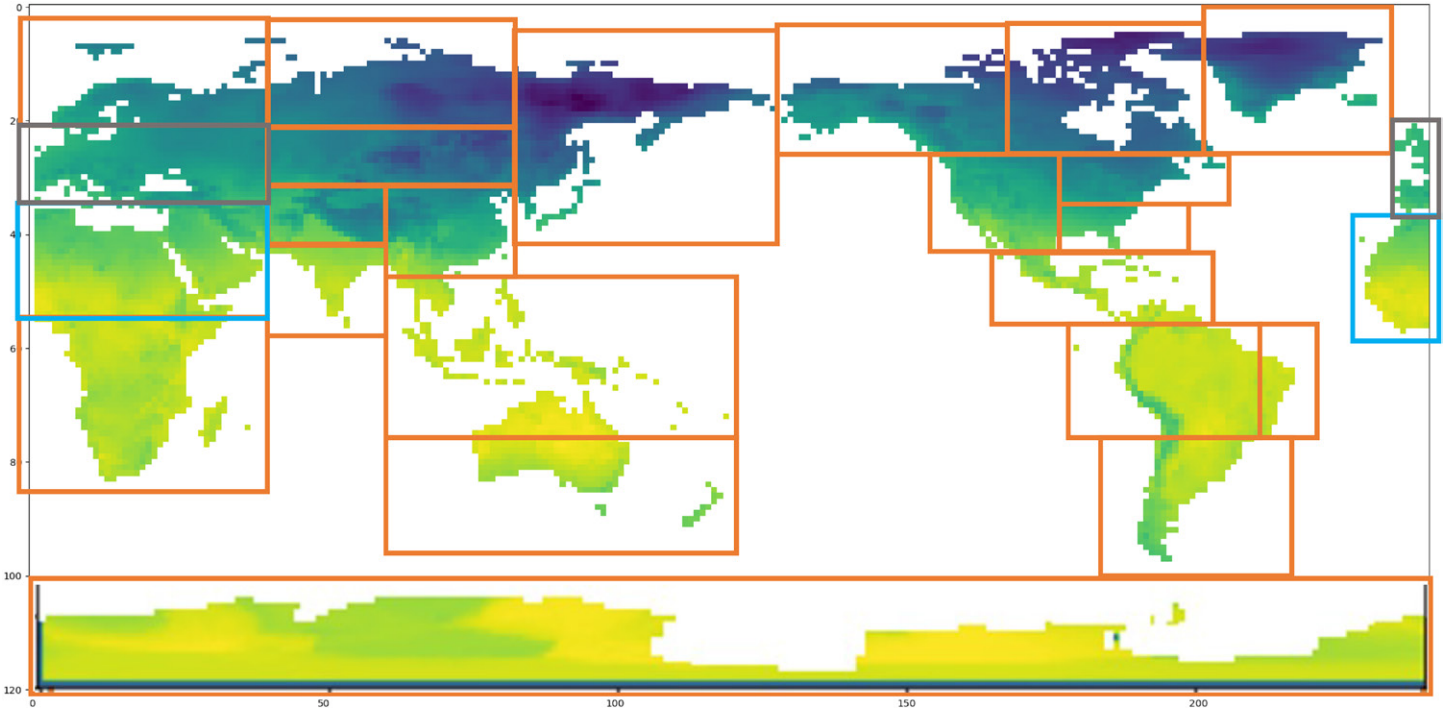

Fig. 4. The 23 regions used by UConn to build separate models. The shading shows the observed temperature on a random day.

The winning entry used AI methods to postprocess the outputs of dynamical models and compute the optimal weights of a multimodel combination. It was the only entry using multiple S2S models, consistent with previous findings that multimodel combinations outperform individual S2S models in general (Vigaud et al. 2017, 2019).

This competition entrained a large number of participants, but only few of them were able to provide skillful S2S forecasts (superior to climatological and ECMWF forecasts), and the positive skill levels of the winning entries remain modest, especially for precipitation, in line with the challenging nature of S2S prediction (Robertson and Vitart 2019). More conventional regression methods were also included by CRIMS2S and BSC, and more work is required to demonstrate that modern machine learning methods like random forest classification and CNNs can significantly outperform these. In addition, the 1-yr test period might be too short to properly evaluate these methods. Therefore, this competition is only a first step in assessing the benefit of AI/ML methods for S2S prediction. There is scope to improve the S2S forecasts even further, by, for instance, including more models or variables.

To stimulate further AI developments, the data tools and software framework on Renku and EWC, including quick-start Jupyter Notebooks developed for this competition, will be maintained, up to at least 2024. These tools allow new AI methods to be easily tested on the same datasets and provide a clean comparison of performance with the winning entries. A second phase of this competition might be envisaged at a later stage.

Acknowledgments. The authors thank the Swiss Data Science Center (SDSC) and the European Centre for Medium-Range Weather Forecasts (ECMWF) for their support to this competition. The CRIMS2S team acknowledges support from the Ministère de l'économie, innovation et exportation (MEIE) of Gouvernement du Québec. The UConn team would like to acknowledge contributions from Krishna Pattipati and Peter Willett from the University of Connecticut, Jason Nachamkin from the Naval Research Laboratory Marine Meteorology Division, and Paolo Braca and Leonardo Millefiori from the NATO STO CMRE. The authors thank the three anonymous reviewers for their suggestions and comments that helped improve the manuscript.

References

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. W. Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04110, https://doi.org/10.1029/2007JD009132.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987, https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132.

He, K., S. Zhang, S. Ren, and J. Sun, 2016: Deep residual learning for image recognition. *2016 IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, IEEE, 770–778, https://doi.org/10.1109/CVPR.2016.90.

James, G., D. Witten, T. Hastie, and R. Tibshirani, 2013: *An Introduction to Statistical Learning: With Applications in R.* Springer Texts in Statistics, Vol. 112, 426 pp., Springer, https://doi.org/10.1007/978-1-4614-7138-7.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Pedregosa, F., and Coauthors, 2012: Scikit-learn: Machine learning in Python. arXiv, 1201.0490v4, https://doi.org/10.48550/ARXIV.1201.0490.

Robertson, A. W., and F. Vitart, 2019: *The Gap Between Weather and Climate Forecasting: Sub-Seasonal to Seasonal Prediction.* Elsevier, 569 pp.

——, ——, and S. J. Camargo, 2020: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *J. Geophys. Res. Atmos.*, **125**, e2018JD029375, https://doi.org/10.1029/2018JD029375.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

——, M. K. Tippett, J. Yuan, A. W. Robertson, and N. Acharya, 2019: Probabilistic skill of subseasonal surface temperature forecasts over North America. *Wea. Forecasting*, **34**, 1789–1806, https://doi.org/10.1175/WAF-D-19-0117.1.

Vitart, F., and Coauthors, 2015: Sub-seasonal to seasonal prediction: Linking weather and climate. *Seamless Prediction of the Earth System: From Minutes to Months*, G. Brunet, S. Jones, and P. M. Ruti, Eds., 385–401.

——, and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Weyn, J. A., and Coauthors, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002502, https://doi.org/10.1029/2021MS002502.