Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC3
https://doi.org/10.5194/gmd-2022-146-AC3, 2022

Reply on RC2

Chuhan Lu et al.

Author comment on "Data-driven Global Subseasonal Forecast Model (GSFM v1.0) for intraseasonal oscillation components" by Chuhan Lu et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-146-AC3, 2022

General comments:

It is good to see more work being done in machine learning applications for weather and climate. This work aims to build on the work done by Rasp et al (2020) by applying SE-Resnet instead of Resnet. The focus here is on learning the intra-seasonal oscillations components. The physical model being compared to is CFSv2.

Response: Thanks a lot for your encouraging and suggestive comments. We have improved our manuscript as you suggested. Please see details in the revised version of the MS and the responses as follows.

General scientific comments:

1. I have some concerns about the methodology being employed in this paper. One of the first things lacking in the paper is a detailed description of data processing. With data-driven methods it is really important to be transparent about how the data were processed and applied. The reader is referred to WeatherBench paper, but the authors of this paper must outline what steps they took ensure their machine learning models were trained properly e.g. training-validation split, data standardisation/normalisation, what were the inputs for the neural network, how were the inputs structured? All these details are not only important for understanding the performance of the model, but also reproducibility.

Response: Thank you for your suggestions. The data used in this paper are listed as follow. (1) Model training data are provided by the WeatherBench challenge. A detailed description can be found in studies of Rasp et al. (2020), and you can get the latest data set on https://github.com/pangeo-data/WeatherBench. The data set mainly contains ERA5 data from 1979 to 2018, and the horizontal resolution of the dataset used in this paper is 5.625°×5.625°. (2) The forecast results of Z500 and T850 in the CFSv2 model data set for the next 30 days. This dataset was downloaded from https://www.ncei.noaa.gov containing data in 2000-2018.

The original CFSv2 forecast data we download covers the global area with a resolution of 1°×1°. The lead time we used ranges from 1-30 day. We interpolated the CFSv2 data into

the same grid points as ERA5 data in this paper (5.625°×5.625°).

The inputs are geopotential, temperature, zonal and meridional wind at seven vertical levels (50, 250, 500, 600, 700, 850 and 925 hPa), 2-meter temperature, and finally three constant fields: the land-sea mask, orography and the latitude at each grid point. All fields were normalized by subtracting the mean and dividing by the standard deviation. In the training process, we first extract a batch of data from the dataset and stack up different variables into a 8 (batch size) * 32 (latitudinal grid number) * 64 (longitudinal grid number) * 32 (4*7+4) array.

The training set includes ERA5 data in 1980-2015, the validation set in 2016 and test set in 2017-2018.

Accordingly, we have added the description of about how we get and process the CFSv2 dataset as well as ERA5 data set in more detail in the Sec. 2.1 (data and method) of our new MS.

2. As the field of machine learning is an empirical science, it would be good to see the reasons behind why certain architectures were chosen e.g. number of residual blocks was set at 17 with two convolutional blocks -- why is this the best setup?

Response: Thank you for your precious advice. The forecast model used in this paper is developed based on the ResNet model designed by Rasp et al. (2021). In their study, it was found that the ResNet structure performed well for the prediction of Z500 and T850, so we used the same residual block structure with two convolutional blocks. As for the number of residual blocks, we tested networks with different numbers of blocks and found that compared with the 19 residual blocks of the original ResNet model in Rasp et al. (2020), reducing the number of residual blocks can also get comparable or better prediction effect, so number of residual blocks was set at 17 in our first MS. However, for continuous forecast, we determined the number of residual blocks of SE-ResNet model as 25 through experiments. The comparison of the networks' performance with different numbers of blocks is shown in the figure below.

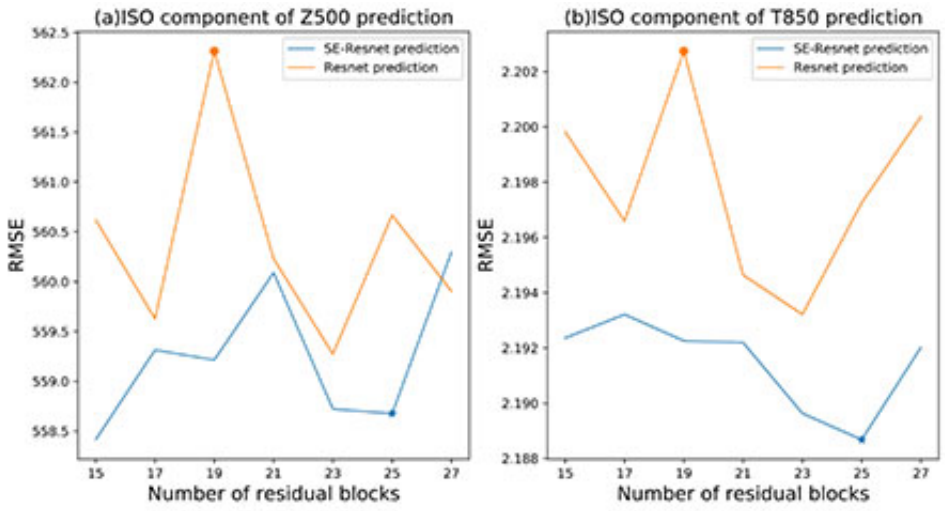

Figure. S1 Comparison of the average RMSE for 10-30 day of (a) Z500 and (b) T850 achieved by training continuous model using different numbers of residual blocks (o and * represent the number of residual blocks used by ResNet and SE-ResNet model in this paper, respectively)

In addition, since there are different factors in different levels entering into the network, it is natural for us to think that the contribution made by different factors to the final result might be different as some meteorological elements are more tightly related to each other than others. In fact, according to Rasp et al. (2021), when predicting T850, the factor that contributes the most to the final is Z250, significantly larger than other factors, that means more attention should be paid to this kind of factors. That is why we decide to use the SE block to automatically select out the more important factors.

3. I am also concerned by point-to-point prediction setup with 30 models for different lead times. This suggests the models are not generalising and being optimised for a very specific subset of the problem. I would suggest asking the question, what is the real-world application of this? Training a separate model for each day of lead time is not efficient and suggests the models are not learning any of the underlying physical processes and therefore require retraining every time there is an improvement in the training model data.

Response: Thank you for critical your advice. In order to learn the underlying physical processes, we have changed the original direct model into a continuous model (both defined in Rasp et al. (2021)). We have tested these two types of models and found out that the performance of the new model is comparable with the old models. Particularly, the average RMSE for 10-30 day of direct model is 552.48 m^2 s^{-2}(2.17 K) and average RMSE for 10-30 day of continuous model is 558.68 m^2 s^{-2}(2.19 K).

4. There are extensive comparisons between CFSv2, Resnet, SE-Resnet, Climatology error and accuracy, which could benefit from having a table to make it easier to interpret, but some of the differences are so small that I would question its statistical significance. The main objective of the paper appears to be to show how deep learning can improve on the physical model's sub-seasonal forecast. However, for that comparison to be fair, it is important to show how well CFSv2 does under different setups and against other physical models. My question would be why choose CFSv2 for this comparison? Is that best model there is for these lead times? Also, the comparisons between SE-Resnet and Resnet do not appear to be statistically significant. In order to understand how different parameters impact ML model's performance, a Pareto front type analysis is required. For understanding the impact of random noise on the ML model, cross-validation should be used to show the difference between two different architectures is consistent and statistically significant.

Response: The CFSv2 model we choose is widely-used around the world for extended-range forecast and represents the top ability in weather forecasting. Also, considering we are extracting the low frequency component as the predictand, the filtering method prefers input data being continuous time series. So we also chose the CMA, UKMO, KMA and NCEP model outputs forecast by S2S (sub-seasonal to seasonal prediction project) from ECMWF, and they are the only four models that provide daily extended-range prediction in consistent with our models as well as CFSv2. You may note that the forecast we added all have a weaker performance compared with CFSv2. That is perhaps because according to Saha et al. (2013), the final result of CFSv2 operational forecast is actually an ensemble mean of 16 runs, but the forecasts we added later only contains one control run. So that is the reason why the gap between CFSv2 forecast and other forecasts are that wide (Fig. S2). It can also tell the superiority of CFSv2.

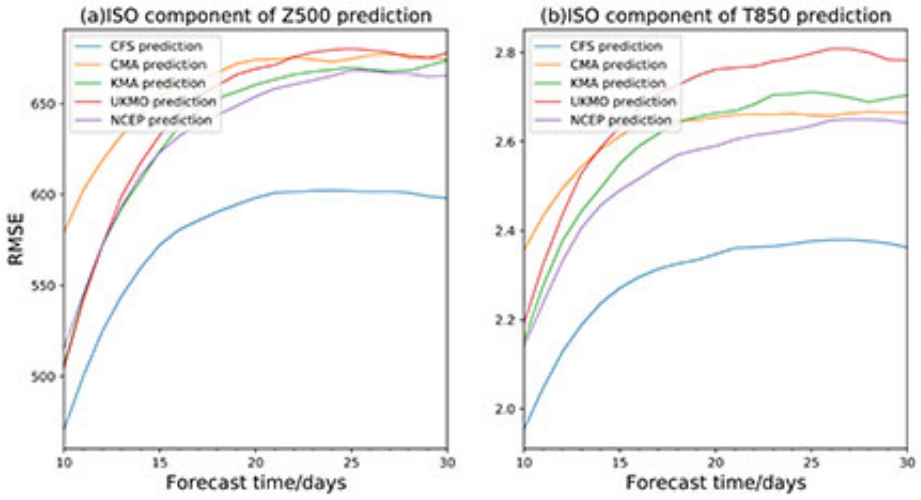

Figure. S2 Comparison of the RMSE for 10-30 day of (a) Z500 and (b) T850 achieved by different models

As for the question that how different parameters impact the model's performance. We carried out some experiments by training SE-Resnet and Resnet with different numbers of residual blocks. As is shown in above Fig. S1, we are able to know how the parameter of residual block number influence the performance of the models, by comparing the average RMSE these models have in forecasting Z500 and T850 10-30 day ahead. Finally, we chose the 25 blocks, which is the best choice due to the lowest RMSE for both Z500 and T850.

We also used different test sets to verify the structural differences between SE-ResNet and ResNet models (Tab. S1) for the cross-validation. The results show that the prediction effect of SE-ResNet is stably better than that of ResNet as the test set changed. Particularly, the average gap for Z500 is 4.03 $m^2 s^{-2}$, the average gap for T850 is 0.01 K. The table shown below not only indicate that the addition of SE module can improve the ResNet model to a certain extent, but also shows that the superiority is steady as for all the test subsets, SE-Resnet outperforms ResNet. Based on these facts, we can say that the SE blocks significantly improve the performance of the network.

Table. S1 Comparison of the average RMSE for 10-30 day of Z500 and T850 achieved by SE-ResNet and ResNet model (SE-ResNet/ ResNet) using different test year

Test year	Z500[$m^2 s^{-2}$]	T850[K]
(2007,2008)	561.14/565.21	2.19/2.20
(2009,2010)	567.92/572.66	2.24/2.26
(2011,2012)	570.70/574.64	2.18/2.20
(2013,2014)	556.41/560.09	2.20/2.21
(2015,2016)	567.05/571.18	2.19/2.20
(2017,2018)	558.68/562.31	2.19/2.20

Specific comments:

5. Figures 2, 7, A1 need to be bigger because in their current size it is almost impossible to see the details being discussed. Perhaps this was just formatting issue for the pre-print.

Response: Thank you for your suggestions. We have modified the size of these pictures.

6. Line 246: 1.01% lower RMSE is a very small difference. It would be good to see statistical significance. Please see above my suggestions of Pareto front and cross-validation.

Response: To verify the difference between them, we conducted average test on the prediction effects of ResNet and SE-ResNet model. The results showed that the two models passed the significance test (α = 0.05) within 10-16 day of continuous forecast (Fig. S3). However, with the extension of the forecast time, the prediction effects of the two models are close to the climatology prediction, and the difference is gradually reduced, which leads to a smaller average increase of SE-ResNet model within 10-30 day.

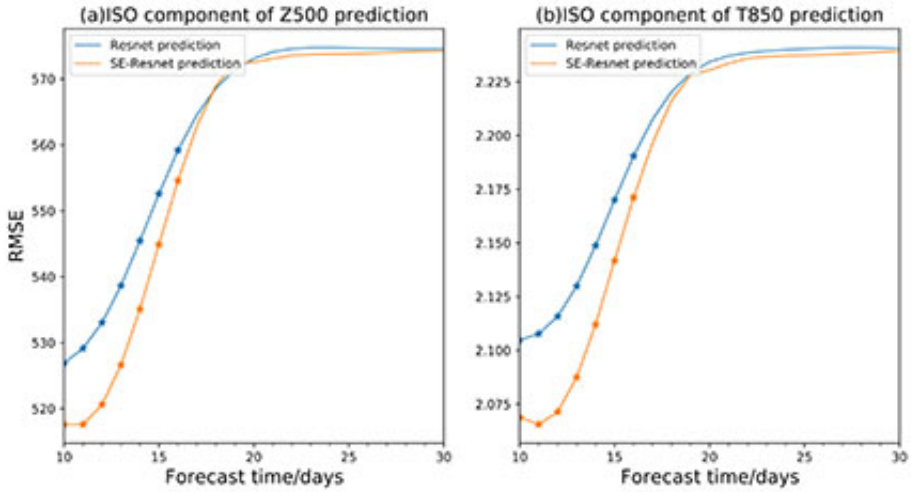

Figure. S3 Comparison of the RMSE for 10-30 day of (a) Z500 and (b) T850 achieved by ResNet and SE-ResNet model (* represents that the corresponding forecast time passes the significance test (α = 0.05))

7. Lines 252-254: There is some odd phrasing which talks about 50% of deep learning models forecast being below climatology, but for CFSv2 50% of forecasts are above climatology. I am sure the authors did not mean to suggest this, but it currently reads as if CFSv2 and deep learning models are the same, but phrased in favour of deep learning models.

Response: Thank you for your advice. What we want to express is that 75 % of the samples predicted by the deep learning model are below the climatological forecast in 10-15 day, and more than 50 % of the samples predicted remain below the climatological forecast after that. While the CFSv2 model predictions have more than 50 % of the samples higher than the climatological forecast in 16-20 day and beyond. So the deep learning models outperforms CFSv2. We have added the new description in our new MS.

8. Line 265: 0.75% RMSE improvement. Statistical significance?

Response: Thank you for your suggestion. This question is the same one posed on line 246, please refer to the answer above.

9. Line 272: It would be good to see the RMSE values for other models quoted too. A table of comparison would be really useful for at-a-glance interpretation.

Response: Thank you for your advice. We have also provided a table version of the RMSE

comparisons. Also, we have add three new forecast models into the comparison, the CMA model, the UKMO model, the KMA model and the NCEP model. Please refer to Tab. S2 in our new MS.

Table. S2 Comparison of the average RMSE for 10-30 day of Z500 and T850 achieved by different model

Model	Z500[m^2 s^{-2}]	T850[K]
CMA	658.19	2.61
UKMO	644.40	2.68
KMA	637.31	2.59
NCEP	634.47	2.53
CFS	577.65	2.29
ResNet	562.31	2.20
SE-ResNet	558.68	2.19

10. Line 288, 291: SE-Resnet and Resnet are so close that there appears to be no difference between them. What is the main benefit of SE-Resnet for this application?

Response: Although there is only a slight improvement between SE-Resnet and Resnet in RMSE and only passed the significance test ($\alpha = 0.05$) within 10-16 day of continuous forecast, we believe that the SE block plays an indispensable role here. After introducing this block into the network, we are able to determine the importance of the input variables so that the model's simulation of the occurrence and development of the weather system is more realistic, as is shown in Fig. 7 of MS.

11. Line 315: SE-Resnet and CFSv2 are very close, so what are the improvements being brought about by SE-Resnet?

Response: The final results of the two networks are close but there are actually significant improvements in SE-Resnet. We performed statistical tests for the RMSE of the models and found out that the numbers passed the test ($\alpha = 0.05$) within 10-30 day, and the RMSE of SE-ResNet is 558.68 m^2 s^{-2}(2.19 K) while the RMSE of CFSv2 is 577.65 m^2 s^{-2}(2.29 K), which is able to prove that significant improvement exists.

Please also note the supplement to this comment:
https://gmd.copernicus.org/preprints/gmd-2022-146/gmd-2022-146-AC3-supplement.pdf