Dear Reviewer,

Thanks for your comments on our paper. Detailed comments and responses are as follows.

The paper applies a machine learning technique for downscaling and calibration of precipitation based on remotely sensed inputs that also aims to incorporate the spatial structure of rainfall using spatial autocorrelation. The idea of paper is interesting and it also has a organized structure which is generally well-written. However, based on the methods applied and discussion of the results, the paper has several shortcomings that need to be addressed and further explained prior to publication.

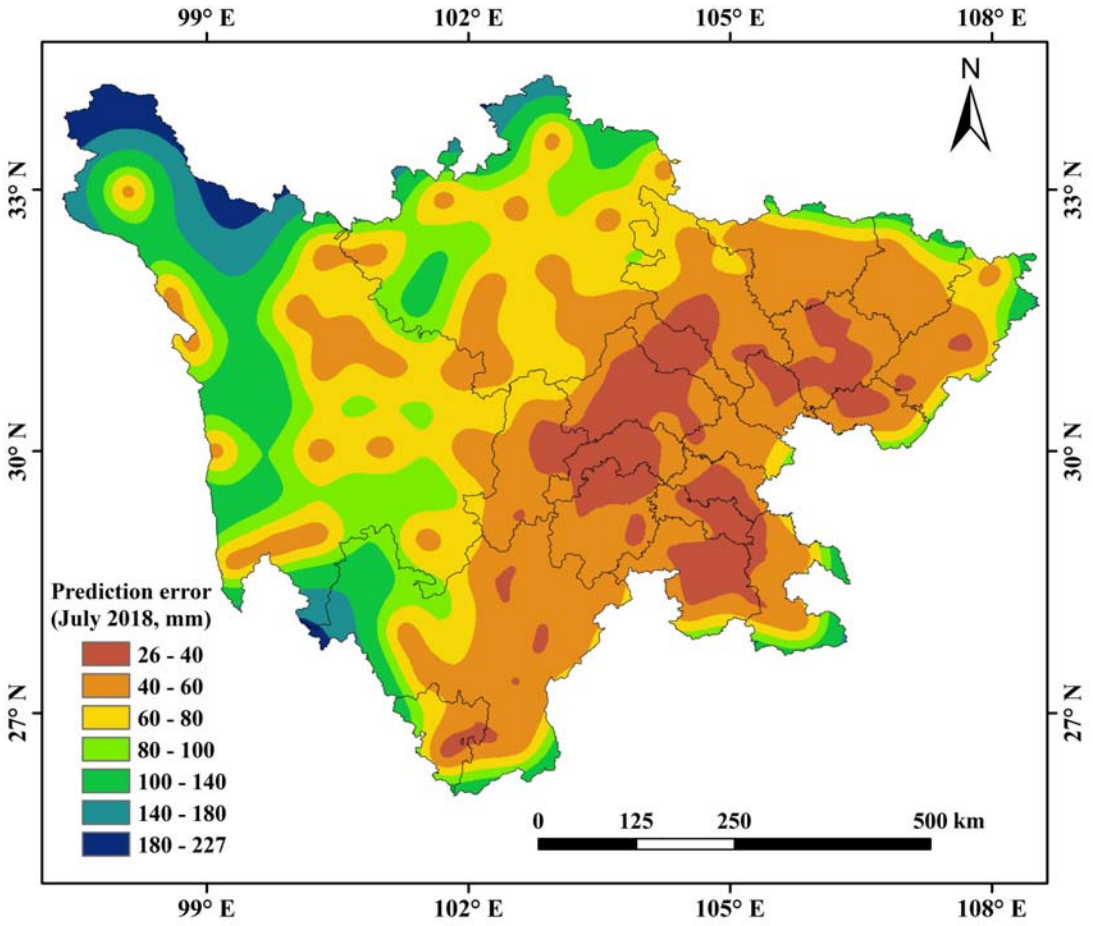
Major comments:

[1] Several aspects of the OK based interpolated maps at 1k and 10k resolutions are not fully convincing. First, the accuracy of the OK-derived maps should be reported in order to determine reliability of the maps. Errors in the interpolated maps are going to be propagated to the errors in the spatial RF model because it is one of the covariates used, so they are important. It would be interesting to see if the large RMSE's in the middle part of the study area in fig.7 also show up with large errors or variance in the OK maps.

[Reply]

Since the wettest month is July 2018 (Fig. 2), it is taken as an example to show the variance of the prediction errors of OK. The prediction error map derived from Eq. (4) shows that the errors in the west are larger than in the east, and in the boundary are

larger than in the inner. It can be inferred that large errors are mainly located in the areas with the sparse distribution of rain gauges, which are not related to the RMSE distribution (Fig. 7) and precipitation (Fig. 8).



(b) Prediction error map

Fig. 9 Semivariogram and prediction error map of kriging on the wettest month (July 2018)

The above information will be shown in the revised paper.

[2] Related to this, the authors also need to further clarify the interpolation of a 1km image based on a 10km IMERG images using OK, which is a raster-to-raster interpolation performed (lines 273-284). A coarse to fine raster-based interpolation seems unusual, so that authors need to further describe this step.

[Reply]

For IMERG interpolation, the raster-based IMMERG were first transformed into point-based form with spatial coordinates (e.g. *x* and *y*) and precipitation values, and then the scattered points were interpolated by OK to produce a map with the given resolution.

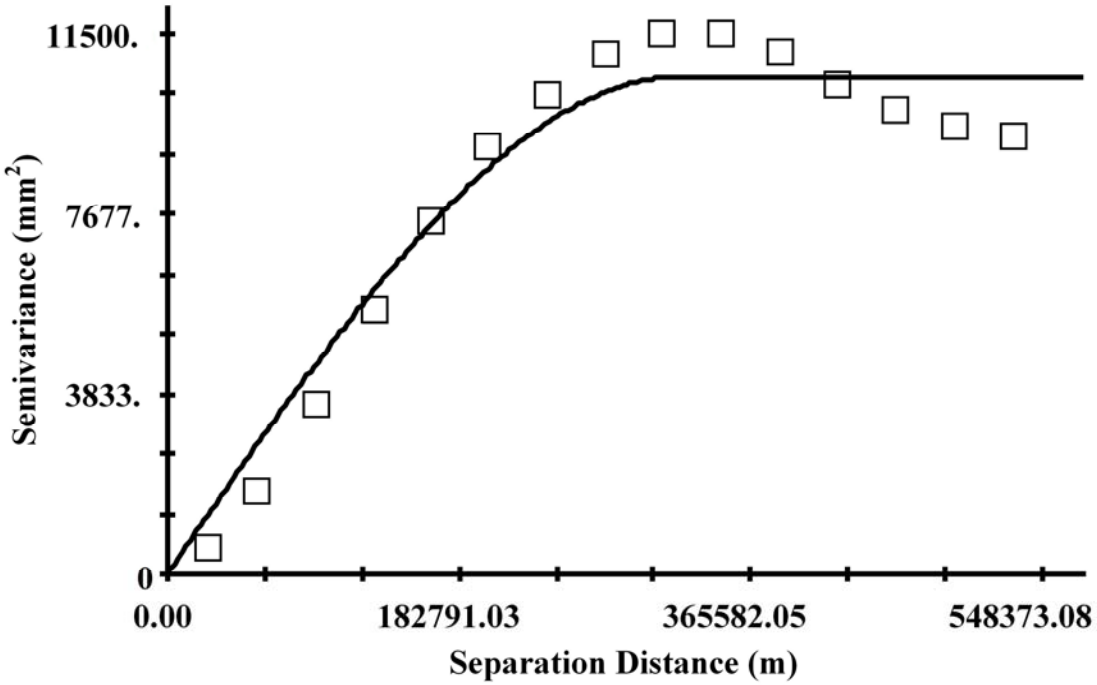
The above information will be added to the revised paper.

[3] The parameters tested and chosen for all the models, including the semi-variogram should be reported otherwise the study is not reproducible.

[Reply]

Kriging was used to produce the 10 km and 1 km satellite-based precipitation products and to interpolate the rain gauge observations for each month of the 5 years. Namely, 3*12*5=180 variograms were used in this study. Similarly, the other methods require at least 12*5=60 groups of parameters. The information is so much that cannot be listed in the paper due to the page limitation.

Since the wettest month is July 2018 (Fig. 2), it is taken as an example to show the semivariogram of OK. For OK, the semivariogram and its prediction error map are shown in Fig. 9. It can be found that it is a spherical model with the nugget variance (C0) of 10.0 m^2 , sill (C0+C) of 10,560 m^2, residual sum of squares (Rss) of 8,800,611 m^2, range (A0) of 321,000 m, and fitting R^2 of 0.962, respectively.



Semivariogram of OK

[4] It appears that the sRF model (also for the other ML techniques applied) did not include a separate testing phase . This is a standard approach applied when assessing the accuracy of a ML methods. I would suggest to also validate the models using an independent test that is not used in the training phase. Or re-configure the ML methods to split the total data into a training and a test set.

[Reply]

All the methods were assessed with separate testing points. The detailed information is as follows:

To quantitatively analyze the performance of all the methods, all rain gauge observations were randomly divided into *l* folds (e.g. *l*=10), where the *l*-1 folds (i.e. training/validation data) was used to construct the model, while the remaining one set

(i.e. testing data) to assess the performance of the model (Xu and Goodacre, 2018). During model construction, the *l*-1 folds were randomly divided into training and validation datasets with the proportions of 80% and 20%, respectively, where the former was used to train the model and the latter to validate the model for tuning parameters. Then, the model with the optimized parameters was assessed using the testing data. The aforementioned process was repeated *l* times until all folds were taken as the testing data.

It can be found that the testing points were not used to construct the model.

[5] Discussion of the results focuses more on the positive aspects of using sRF but the authors do not give a balanced view by providing a critical analysis of the results of sRF. For instance, the accuracy metrics presented highlight that sRF performs well compared to the other models. However, visual comparison of the boxplots of these metrics alone in figs. 8-9 shows comparable accuracies all the models based on their range and median. Significant differences between the accuracies obtained, particularly in relation to sRF, should be reported to provide gravitas on the authors claim that sRF outperforms the other models.

[Reply]

In the revised paper, a balanced view will be added to the revised paper to give a fair assessment on the performance of the proposed method. The information is as follows:

Although SRF-DC shows promising results than the classical methods, it still suffers from some limitations, which should be solved in the further researches.

Firstly, SRF-DC is more complex than Bi-SRF and SRF-GDA, since SRF is used in both downscaling and calibration in SRF-DC. Hence, applying SRF to downscale IMMERG might not be prerequisite since SRF-DC is only slightly better than Bi-SRF. However, SRF should be used to calibrate IMMERG due to the obviously higher accuracy of SRF-DC than SRF-GDA.

Secondly, SRF-DC has an obvious underestimation on high precipitation values mainly due to the omission of some important land surface variables for precipitation estimation. Thus, other available variables such as soil moisture (Fan et al., 2019; Brocca et al., 2019), and meteorological conditions such as cloud properties (Sharifi et al., 2019) should be adopted to further improve IMERG quality.

Thirdly, the correction of satellite-based precipitation on higher-temporal scales (e.g. daily or hourly) is challenging and valuable (Wu et al., 2020; Chen et al., 2020b; R. Lima et al., 2021; Sun and Lan, 2021). Whether SRF-DC could be applied on these scales requires further validation.

Finally, numerous satellite-based precipitation products have been available, and each one has its shortcomings and advantages for the capture of spatial precipitation patterns (Chen et al., 2020c; Baez-Villanueva et al., 2020). Thus, the fusion of multiple precipitation products based on SRF-DC is a promising alternative to improve the quality of precipitation data.

[6] There is an underestimation of precipitation values regardless of the model used based on fig 5. This should be further elaborated in addition to the three accuracy metrics provided, so the bias of the estimates should also be reported. Furthermore,

for very high precipitation values (e.g. >400mm), the scatter of the points in fig.5 becomes larger, indicating that all the models tested perform poorly at v. high rainfall amounts. It could be insightful to assess separately how the models compare for v. high rainfall conditions, since prediction of these extreme cases need to be generally improved.

[Reply]

We will report the bias of the estimates with respect to mean error (ME). Thus, the ME will be added to the scatterplot in the revised paper.

Moreover, we compare the performance of all the methods on the very high precipitation values (e.g. >400mm). To quantitatively analyze the performance of all methods on the observed values greater than 400 mm, their accuracy measures are shown in Table 2. Results illustrate that all methods have poor results for these observations. A possible reason is that high precipitations are often caused by complicated environmental factors, which cannot be sufficiently explained by the constructed predictors-precipitation relationship. In terms of ME, SRF-GDA ranks the first, which is followed by kriging and SRF-DC. However, their ME values are less than -70 mm. With respect to RMSE and MAE, kriging performs the best, which is closely followed by SRF-DC, and with respect to CC, SRF-DC with the value of 0.64 outperforms the others. Overall, considering the poor performance of kriging for mapping spatial precipitation distribution, SRF-DC seems the best choice for the extreme precipitation estimation.

Table 2 Accuracy measures of all methods for the observed values greater than 400

mm

Method	ME (mm)	RMSE (mm)	MAE (mm)	CC
SRF-DC	-105.54	149.80	124.82	0.64
Bi-SRF	-110.96	156.81	130.67	0.60
SRF-GDA	-74.21	150.10	126.02	0.55
SRFdis	-117.31	160.11	137.29	0.61
Kriging	-86.25	146.94	119.53	0.58
RF	-141.53	177.71	150.83	0.61
BPNN	-118.88	171.23	142.00	0.57
GWR	-139.02	178.85	145.19	0.57
IMERG	-136.22	173.24	143.69	0.55

The above information will be shown in the revised paper.

[7] It is unclear how the importance measures are calculated from fig. 13, so this should also be included in the methodology of the paper. Furthermore, discussion of the rankings could be made more in depth by determining whether they agree or deviate (and why they do) from known controls on rainfall distribution.

[Reply]

To measure the importance of the *i*th predictor, its values are permuted while the values of the other predictors remain unchanged. Then, the OOB error based on the permuted samples is computed. Next, the importance score of the *i*th predictor is computed by averaging the difference between the OOB errors before and after the permutation. With the scores, the importance of each variable can be ranked.

Based on RF, the relative importance of each predictor (i.e. predictor importance estimate) is shown in Fig. 10. Results show that precipitation from kriging interpolation has the most importance. This is because the interpolated value is directly related to precipitation. Kriging estimation is followed by the downscaled precipitation. Longitude is the third most important variable, which is followed by latitude. This result is consistent with that of Karbalaye Ghorbanpour et al. (2021). They indicated that compared to NDVI, LST and DEM, longitude ranks the first with respect to importance score.

The three LSTs also have a great impact on the precipitation estimation, where LST_D seems slightly more important than LST_N and LST_{D-N}. NDVI has a slight effect on the precipitation, which ranks last but one. This might be due to the fact that NDVI is influenced by both precipitation and temperature in the study site, and the low temperature above certain elevations hinders the vegetation growth. It should be noted that it is less likely that the response of vegetation to precipitation has the delay in the study site, since SRF-DC on the monthly scale is more accurate than SRFdis on the annual scale.

Among the 12 predictors, aspect has the least importance. This conclusion was also obtained by Ma et al. (2017) for downscaling TMPA 3B43 V7 data over the Tibet Plateau. Compared to aspect, DEM, terrain relief and slope seem more important, since precipitation shows obvious relationships with topography. This is consistent with previous studies (Immerzeel et al., 2009; Jing et al., 2016).

The above information will be shown in the revised paper.

[8] The authors already indicate that there is a delayed response of vegetation to rainfall. It is perhaps expected that the NDVI is one of the least important factors in the sRF model. But actually, this also provides an opportunity to also explore the lagged values of the predictors (and not only NDVI) with known delayed responses to rainfall.

[Reply]

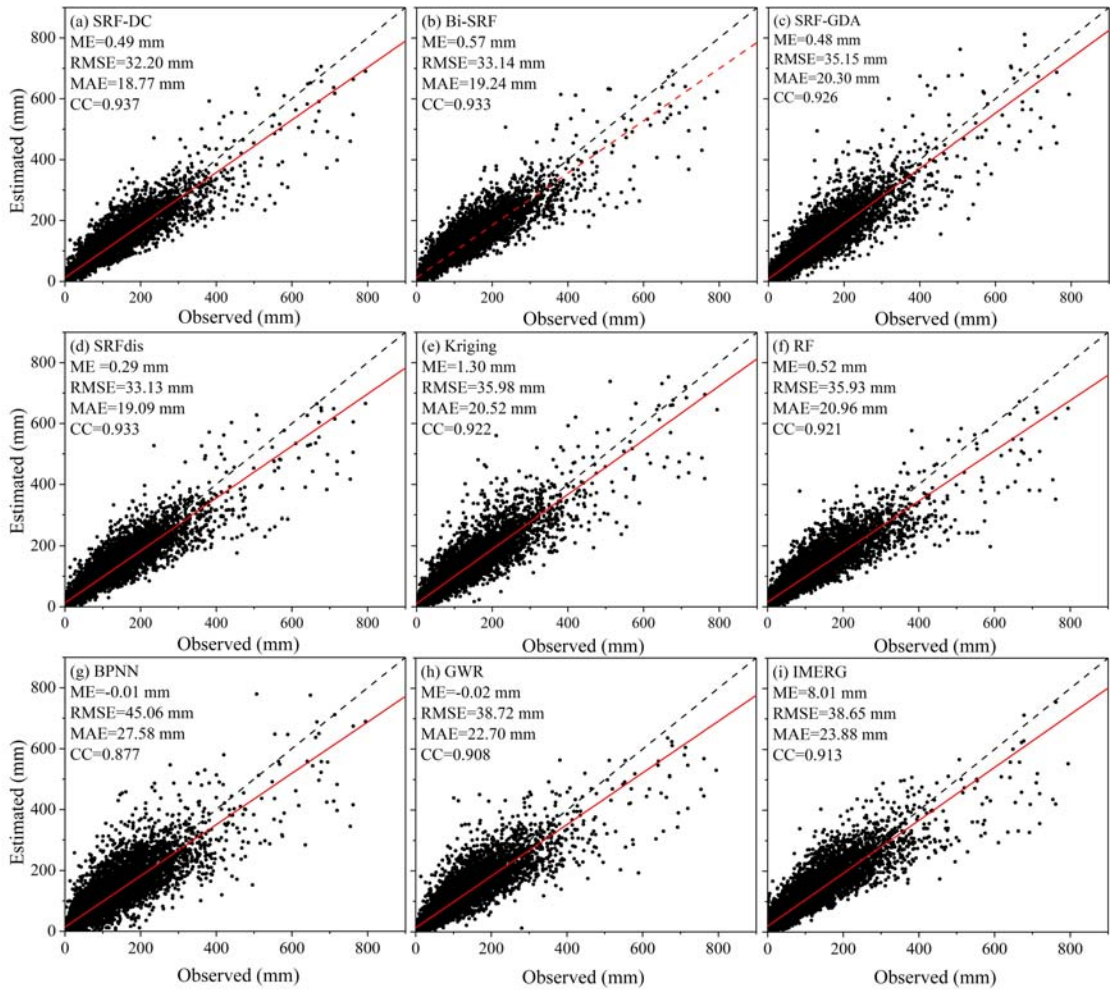
This might be a reason for the least importance of NDVI. However, in this study, it is found that SRF-DC on the monthly scale is slightly more accurate than that on the annual scale (i.e. SRFdis), indicating that the response of vegetation to precipitation has no obvious time delay.

Minor comments:

[9] The captions of the figures need to be improved. Some of the features in multi-plot figures are hard to understand because of the captions are highly simplified.

[Reply]

The captions of the figure will be added to better understand in the revised paper. It is as follows:



[10] The final version of the manuscript will benefit for another round a English check as some sentences a phrased a bit vaguely (e.g. line 150-151)

[Reply]

Our paper will be polished by a naïve English speaker.

Best wishes,

Chuanfa Chen

Baojian Hu

Yanyan Li