Dear Reviewer,

The comments offered have been immensely helpful. We appreciate your insightful comments on our paper. We have responded to every question, indicating exactly how we addressed each concern.

The manuscript aims to produce a long term dataset of monthly 2m temperature over China at high spatial resolution on a 1x1 km grid. While the objective is appealing due to the challenges related to the complex topography and the irregular data availability in the target region, the applied methods show up with significant issues. The major issues are listed subsequently:
Reply：Many thanks for the comments. We apologize for not expressing ourselves clearly. The method we designed generates highly accurate data products. Test results from meteorological observation sites in the field show that our method is robust and repeatable. We have responded to every question to make the expression clearer and more accurate. The point-to-point responses to the comments are listed below.

[1] The introduction discusses advantages and disadvantages of different information sources for the targeted dataset. While strong arguments for point-wise observational data are presented, long term reanalysis data products are not considered despite they provide consistent and spatio-temporally coherent information on the atmospheric state. It is unclear why such data is not considered to provide predictor variables.

Reply: The reanalysis data have some limitations as the predictor variables.

(1) First, the resolution of the reanalysis data is usually low (e.g. the resolution of ERA5 data is 0.25°). Since the spatial resolution in our study is 1-km, the reanalysis products cannot provide such fine resolution data.
(2) Second, the time span of the reanalysis data can not meet the study period in our study. The period of ERA5 starts from 1979 (Tang et al., 2020) while the dataset we produced starts from 1951.
(3) Third, the reanalysis data are generated using the station observed data, which have uncertainty per se. As shown in the study of Tang (2020), the accuracy of ERA5 data in China is relatively low. The satellite-based and atmospheric reanalysis precipitation estimates are highly constrained by errors (Yin et al., 2021).
(4) The model designed in our study can generate high-resolution datasets without using the reanalysis data.

Considering the above, we did not consider the reanalysis data in our study.

[2] The method of data splitting leads to strong autocorrelation between the training and test dataset. Due to the spatial proximity of stations in both dataset, a fundamental requirement is hurt, that is the independency (or at least a minimization of dependency) between the training and test dataset. This is especially true for the stations located in the flat eastern parts of China with a dense observational network. Thus, the statistical model are prone to learn nearest neighbor-relations

rather than learning real abstractions from the features, see, e.g. Kleinert et al., 2021 for a more detailed discussion on the requirement of splitting the test and training data temporally when stations are located close to each other.

Reply: Many thanks for your constructive comments. In machine learning, there are two strategies for splitting the training set and testing set. The first is a spatial division which split the data into the training set and testing set on the spatial field. The second is temporal split which splits the data into non-overlapping time periods for training and testing, e.g. the study you mentioned (Kleinert et al., 2021). However, there is no standard method for splitting the training and testing dataset. In our study, we used the first strategy for splitting the data, mainly because the following reasons:
(1) The temporal splitting is not appropriate in our study. In the study of Kleinert (Kleinert et al., 2021), they used all the data from 1 January 1997 to 31 December 2007 as the training dataset while it is not feasible in our study to use all the historical data as training sets. Our object is to generate the long time-series data for each month ranging from 1951 to 2020. Thus we need a testing dataset for each month to evaluate the monthly data. In our study, the spatial splitting method can meet the requirements of our study goal.

(2) In the spatial prediction of the environmental variables, numerous study uses the spatial split (Costache et al., 2020; Band et al., 2020; Mohajane et al., 2021; Kutlug Sahin and Colkesen, 2021; Hijmans et al., 2005; Fick and Hijmans, 2017).

(3) The spatial splitting was completed using the "Subset Features" (Geostatistical Analyst) tool in ArcGIS which divides the original dataset into two parts: one part can be used to construct the model; the other part can be used to compare and validate the output. "Subset Features" is the most rigorous way to assess the quality of an output surface. Several studies have used the Subset Features tool of ArcGIS for splitting training and testing datasets in machine learning modelling (Costache et al., 2020; Band et al., 2020; Mohajane et al., 2021; Kutlug Sahin and Colkesen, 2021).

The Subset Features tool divides the data into two subsets. Subset one will have L features, and subset two will have N - L features (with N being the amount of features in the original dataset). The features are divided by generating random values from a uniform [0,1] distribution. If the random value is less than L/N, the feature is assigned to the first subset. If not, the feature is assigned to the second subset. (source: https://desktop.arcgis.com/en/arcmap/latest/extensions/geostatistical-analyst/how-subset-features-works.htm).

(4) We used the 10-fold cross-validation when training the models (Line 193 in the manuscript).

(5) The spatial distribution of the testing data is similar to the spatial distribution of all the data, which is in conformity with the stratified sampling scheme in the machine learning.
We will discuss the splitting strategies in the discussion part of the manuscript.

[3] Only static features are used as predictors which implies that a model must trained for each month (!) of the period under consideration. Thus, dynamic information on the atmospheric state can exclusively deduced from the optimization procedure on the predictand. It is strongly

recommended to introduce dynamical data as a predictor variable instead. Besides, the chosen predictors have periods with neglectable correlation with respect to the target quantity and important features such as the ambient topography (is the meteorological station located in a valley) is absent (see, e.g., Sha et al., 2020).

Reply: Many thanks for your comments.

(1) In our study, the model was trained for each month. We described the model construction in Lines 193-194. The longitude, latitude and elevation are indeed static factors, but we construct the model for each month, respectively, which can reflect the changes of temperature from month to month.

(2) The remote sensing data such as NDVI, land use change and surface temperature are usually not available before 2000 since our data is from 1951 to 2020. Furthermore, the MODIS data are not available for each month from January 2000 to December 2020. As shown in Figure 1, the percentage of the available MODIS images are low in northeast China and southern areas. So the remote sensing data are not appropriate for generating long-term temperature data in our study.

(3) Furthermore, there is inherent data inaccuracy in the remote sensing data itself, such as the land use data.

(4) As shown in the study of Sha et al. (2020), the orography, as represented by elevation fields can help characterize the spatial heterogeneity of 2-m temperature. The meteorological processes are locally embedded with small-scale terrain features. The terrain features including plain, slope, peak and valley are recognized as the semantic contents of terrain. They used the elevation data to represent those terrain semantics. In our study, we used the DEM data as the predictor in the machine learning model. As said in the study of Sha et al. (2020), the terrain semantics can be learned from gridded elevation inputs.

(5) We can obtain the high-resolution dataset using the selected predictors in our study. The accuracy evaluation shows the rationality of the predictors. The model is robust in generating the long-term temperature datasets.

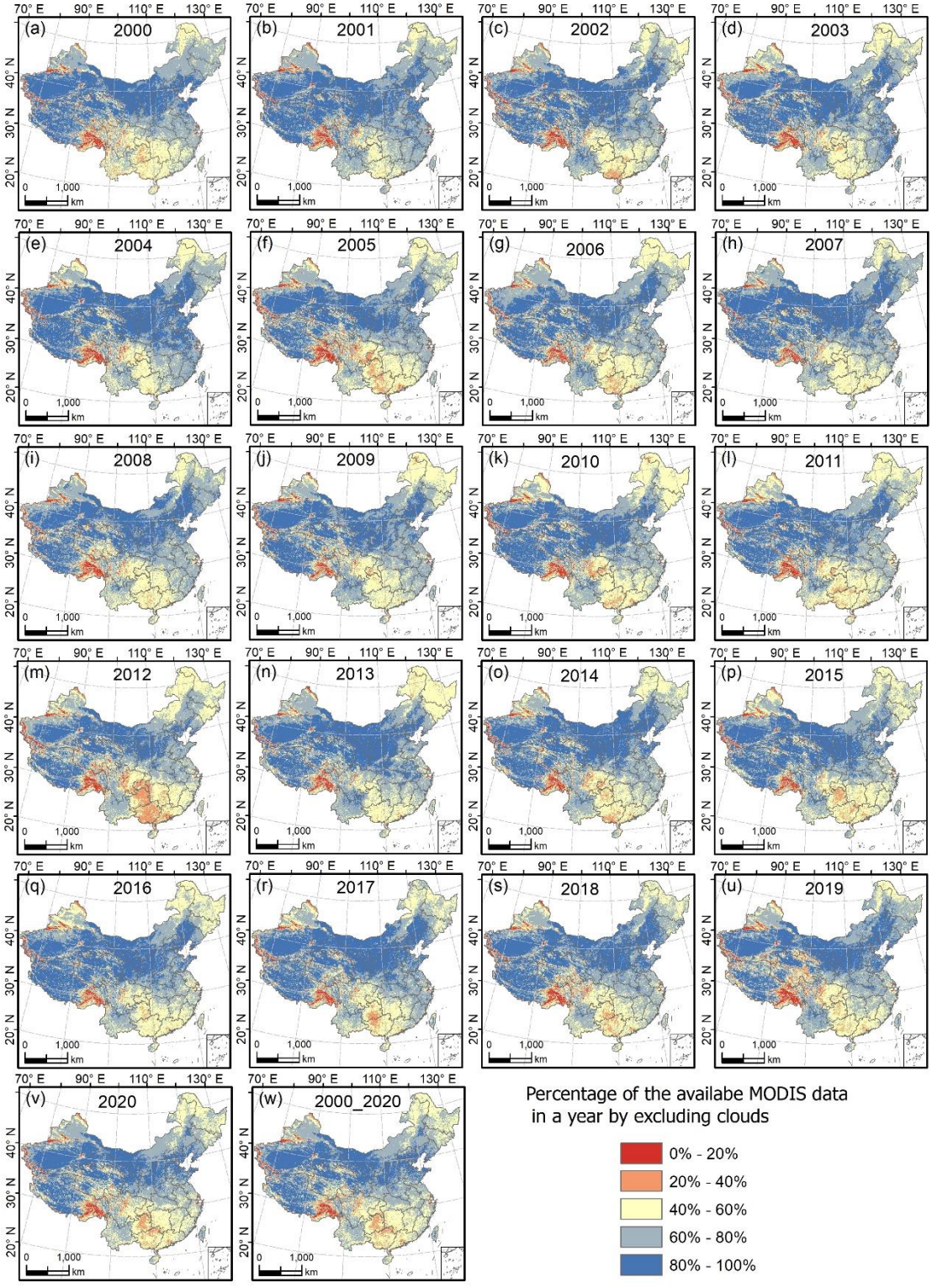

Figure 1 Spatial distribution of the percentage of the available MODIS images in each year (2000 - 2020) by excluding clouds.

[4] The evaluation does not serve the objectives of the study. The stations in the test dataset are dominated by stations over flat terrain with a dense observational network. Thus, potential deficiencies in capturing the variations due to underlying complex topography are hidden. Indeed,

Figure 5 indicates that residuals are considerably larger over the mountainous region.

Reply: We agree with the comment. The meteorological stations in mainland China are unevenly distributed with more stations in the flat terrain and less stations in the mountains. This is the inherent data limitation for modelling continuous raster products in China (Guo et al., 2020; Liu et al., 2018). It is true that the stations in the Qinghai-Tibet plateau are sparse (Xu et al., 2018; Zhang et al., 2016). It is also an existing challenge of the spatial interpolation of temperature using the station data. The potential deficiency in capturing the variations in regions with complex terrain is an existing issue in the current studies. We are working to improve the accuracy of models in complex regions. The altitude information is conducive to the estimation of temperature (Berndt and Haberlandt, 2018).

To show the strength of our data in regions with complicated topography. We took the Tibetan plateau region as an example. We compared the accuracy of our data with Peng's data (Peng et al., 2019) in Tibetan plateau. In our study, the accuracy in the Qinghai-Tibet Plateau is relatively good. We used the mean temperature data of Peng et al. (Peng et al., 2019) to make a comparison. The testing stations which were not used in the model training were used to make the comparison. As shown in Figure 2, the RMSE of most months for GPR is lower than Peng and GPR has smaller variation in RMSE. The R^2 of GPR shows good accuracy from January to December, with smaller variation in each month, while for Peng's data the variation in summer is quite high.

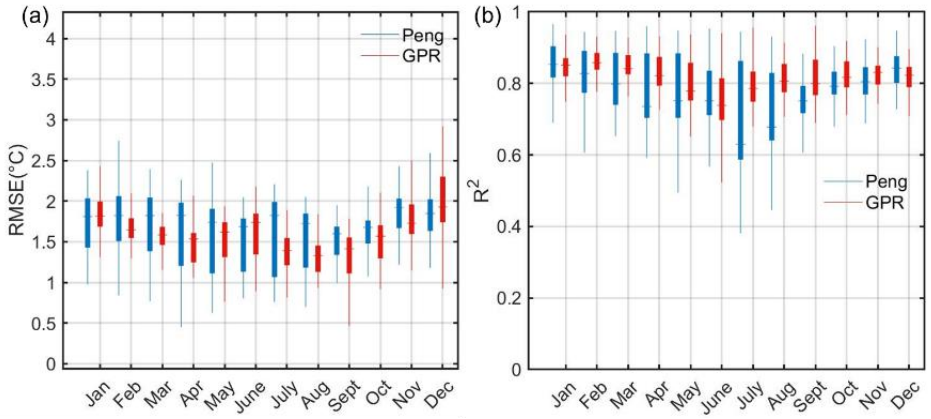

Figure 2 Comparison between the GPR data in our study and the Peng's data in the Tibetan region

[5] Several issues in the follow-up study are present such as (a) a focus on large-scale temperature patterns instead of fine-scale patterns in Section 4.2. to reason the high spatial resolution of the dataset, (b) the interpretation of patterns in the Xinjiang region which look like artefacts (bulls-eye pattern in winter months) and (c) the missing notification on the better performance of the reference method ANUSPLIN for July-months in the 70s, 80s and 90s.

Reply: The spatial resolution of the temperature dataset in our study is 1 km. Since the territory of China is large, it is challenging to produce fine-scale temperature data. Besides, the scale of 1 km is the resolution of a lot of high-resolution datasets for mainland China, like the 1 km monthly temperature and precipitation dataset (Peng et al., 2019), 1 km daily surface air temperature product over mainland China (Chen et al., 2021), a high-resolution crop phenological dataset for three staple

crops in China (Luo et al., 2020). The 1-km resolution is high enough for mainland China which can satisfy a lot of the requirements in other scientific research or practical applications. The 5-km spatial resolution dataset for Spain which is way smaller than China is also treated as the high-resolution dataset (Serrano-Notivoli et al., 2019). We admit that the finer resolution data may provide more detailed information but the 1-km resolution data is high enough for multiple studies.

(b) The bulls-eye pattern in the Xinjiang region in the winter months is induced by the complex topography, which just shows that the model captured the detailed local differentiation of temperature due to topographic conditions. As shown in Figure 3, the region which has a relatively lower temperature (Section 4.2 Figure 6) is because of the high altitude.

(c) Considering the proven power of ANUSPLIN in predicting meteorological variables, the GPR yields relatively satisfactory results. The accuracy of ANUSPLIN for July-months in the 70s, 80s and 90s is slightly higher than GPR while the accuracy of GPR still performs relatively well. ANUSPLIN uses the thin-plate smoothing spline algorithm which allows the introduction of multivariate linear sub-models with complex model coefficients to be calculated. Using the same computational resources, ANUSPLIN is more time-consuming than GPR. Besides, GPR has higher accuracy in the winter months. We will discuss more in the discussion section.

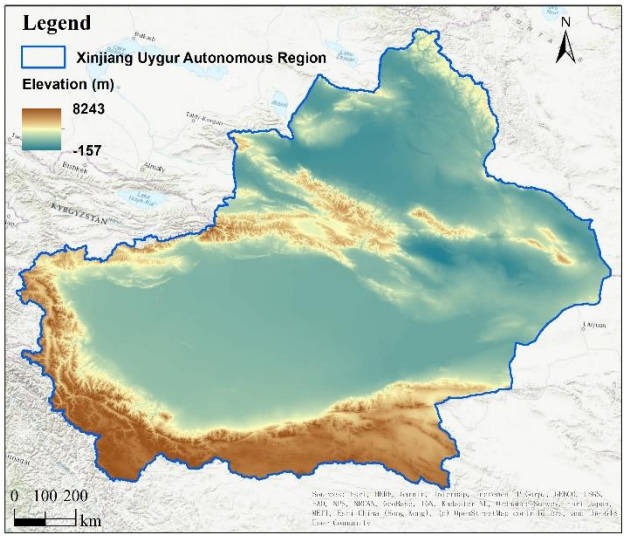

Figure 3 The elevation of Xinjiang Uygur Autonomous Region

[6] The comparison to the competing datasets ERA5 and FLADS is misleading due to the much coarser spatial resolution of these two datasets. A fair comparison would consult datasets with similar spatial resolution such as the dataset described in Peng et al., 2019.

Reply: Many thanks for your comments. We use three datasets ERA5, FLDAS and TerraClimate. The spatial resolution of the three datasets is 27830 meters, 11132 meters, and 4638.3 meters, respectively. Our datasets are 1-km. We resampled all the data to 27830 meters to keep the resolution consistent and then we comparisons. As shown in Figure 4, GPR still outperformed other products.

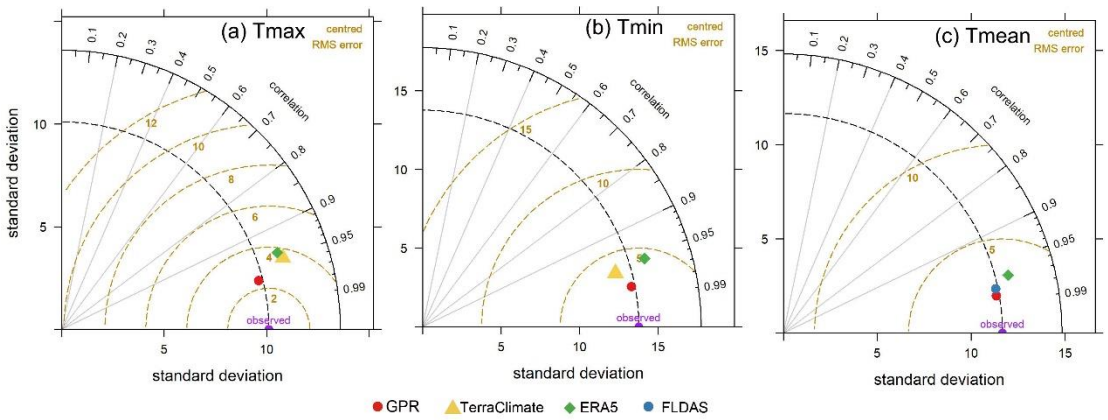

Figure 4 Taylor diagrams displaying a statistical comparison with observations between our products generated using the GPR model and the other products under the same spatial resolution.

Besides, we also compared our datasets with Peng's data as you suggested. As shown in Figure 5, our datasets have relatively higher accuracy than Peng's data on the whole, especially in warm months.

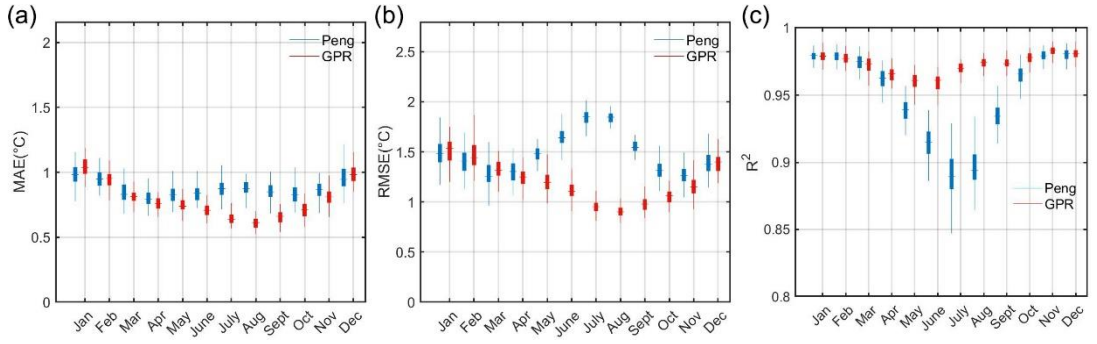

Figure 5 Accuracy comparison between the GPR data and the Peng's data for mean temperature

Further minor issues are:

* Splitting into three distinct datasets is unnecessary. Rather merge it to one dataset with one DOI.
Reply: The Zenodo database has limitations for the data size (max 50 GB per dataset). The zip format file of each dataset is about 30 GB, so we uploaded mean, maximum and minimum temperature data, respectively.

* Refer to statistical and dynamical downscaling techniques in the introduction.
Reply: Thanks for your suggestion. The downscaling technique uses the existing coarse product to produce the high-resolution dataset. The interpolation uses the meteorological station data to generate the spatial continuous grid dataset. The two strategies use different source data, but they have the same objectives. In fact, there are multiple low spatial resolution datasets, such as the Climatic Research Unit (CRU) (Harris et al., 2014), the Global Precipitation Climatology Centre (GPCC) (Schneider et al., 2014; Becker et al., 2013), and Willmott & Matsuura (W&M) (Matsuura and Willmott, 2012) which are generated using the data from the observational stations. It is a reliable way to produce continuous datasets using the observed station data (Peng et al.,

2019).

* Provide references to the problems related to remote sensing data (see l.66).
Reply: Thanks for your reminder. The references are listed below:
(Dong and Xiao, 2016)
(Xiao et al., 2018, p.2013–2016)
(Mao et al., 2019)

* Describe the remapping of the STRM DEM data onto the 1x1 km grid (should be an averaging method).
Reply: We used GEE to export the STRM DEM data as 1*1 km grid. The "Scale" parameter was used to specify the output resolution to 1 km. The concept of "Scale" is illustrated in Figure 6 (https://developers.google.com/earth-engine/guides/scale). The default method of resampling is the nearest neighbour (https://developers.google.com/earth-engine/guides/scale#image-pyramids).

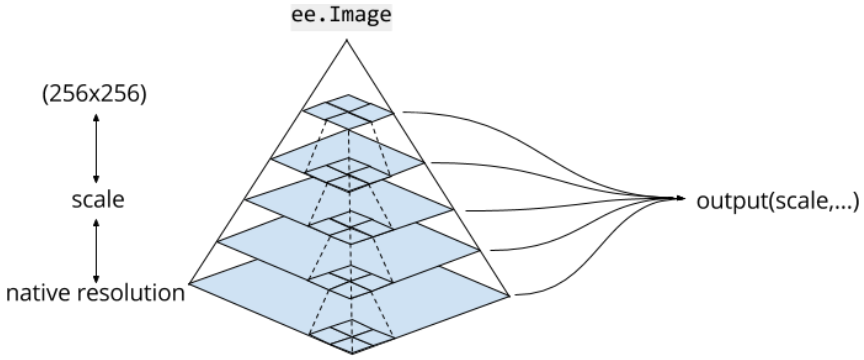

Figure 6 A graphic representation of an image dataset in Earth Engine. Dashed lines represent the pyramiding policy for aggregating 2x2 blocks of 4 pixels. Earth Engine uses the scale specified by the output to determine the appropriate level of the image pyramid to use as input.

* The used software tool MATLAB should be only mentioned once rather than being repeated three times. More details on the respective ML-technique would be appreciated.
Reply: Thanks a lot for your advice. We should mention the MATLAB once. In the light of the limitation of the words in the manuscript, we did not provide so many details for all the machine learning methods but we provided the references or related links which have detailed descriptions of the machine learning methods. In the next revision, we will add more key information about the machine learning methods in the manuscript.

* l.149: Should be 'ensemble machine learning'
Reply: You are right. Thanks for pointing this out and sorry for the wrong spelling.

* l.219: "Unnecessary reference to Equation 4 which directly follows the sentence.
Reply: Thanks for your comment. The references should be removed.

* l.343f. This is sentence is barely comprehensible.
Reply: What we want to express is that the accuracy measures fluctuate with the seasons.

References

Subset Features (Geostatistical Analyst)—ArcGIS Pro | Documentation: https://pro.arcgis.com/en/pro-app/latest/tool-reference/geostatistical-analyst/subset-features.htm, last access: 28 January 2022.

Band, S. S., Janizadeh, S., Chandra Pal, S., Saha, A., Chakrabortty, R., Melesse, A. M., and Mosavi, A.: Flash Flood Susceptibility Modeling Using New Approaches of Hybrid and Ensemble Tree-Based Machine Learning Algorithms, 12, 3568, https://doi.org/10.3390/rs12213568, 2020.

Becker, A., Finger, P., and Meyer-Christo, A.: A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present, 29, https://doi.org/10.5194/essd-5-71-2013, 2013.

Berndt, C. and Haberlandt, U.: Spatial interpolation of climate variables in Northern Germany—Influence of temporal resolution and network density, Journal of Hydrology: Regional Studies, 15, 184–202, https://doi.org/10.1016/j.ejrh.2018.02.002, 2018.

Chen, Y., Liang, S., Ma, H., Li, B., He, T., and Wang, Q.: An all-sky 1 km daily surface air temperature product over mainlandChina for 2003–2019 from MODIS and ancillary data, Data, Algorithms, and Models, https://doi.org/10.5194/essd-2021-31, 2021.

Costache, R., Pham, Q. B., Sharifi, E., Linh, N. T. T., Abba, S. I., Vojtek, M., Vojteková, J., Nhi, P. T. T., and Khoi, D. N.: Flash-Flood Susceptibility Assessment Using Multi-Criteria Decision Making and Machine Learning Supported by Remote Sensing and GIS Techniques, 12, 106, https://doi.org/10.3390/rs12010106, 2020.

Dong, J. and Xiao, X.: Evolution of regional to global paddy rice mapping methods: A review, ISPRS Journal of Photogrammetry and Remote Sensing, 119, 214–227, https://doi.org/10.1016/j.isprsjprs.2016.05.010, 2016.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, 37, 4302–4315, https://doi.org/10.1002/joc.5086, 2017.

Guo, B., Zhang, J., Meng, X., Xu, T., and Song, Y.: Long-term spatio-temporal precipitation variations in China with precipitation surface interpolated by ANUSPLIN, Sci Rep, 10, 81, https://doi.org/10.1038/s41598-019-57078-3, 2020.

Harris, I., Jones, P. d., Osborn, T. j., and Lister, D. h.: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset, 34, 623–642, https://doi.org/10.1002/joc.3711, 2014.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas, Int. J. Climatol., 25, 1965–1978, https://doi.org/10.1002/joc.1276, 2005.

Kleinert, F., Leufen, L. H., and Schultz, M. G.: IntelliO3-ts v1.0: a neural network approach to predict near-surface ozone concentrations in Germany, Geosci. Model Dev., 14, 1–25, https://doi.org/10.5194/gmd-14-1-2021, 2021.

Kutlug Sahin, E. and Colkesen, I.: Performance analysis of advanced decision tree-based ensemble learning algorithms for landslide susceptibility mapping, 36, 1253–1275, https://doi.org/10.1080/10106049.2019.1641560, 2021.

Liu, Z., Liu, Y., Wang, S., Yang, X., Wang, L., Baig, M. H. A., Chi, W., and Wang, Z.: Evaluation of

Spatial and Temporal Performances of ERA-Interim Precipitation and Temperature in Mainland China, 31, 4347–4365, https://doi.org/10.1175/JCLI-D-17-0212.1, 2018.

Luo, Y., Zhang, Z., Chen, Y., Li, Z., and Tao, F.: ChinaCropPhen1km: a high-resolution crop phenological dataset for three staple crops in China during 2000–2015 based on leaf area index (LAI) products, 12, 197–214, https://doi.org/10.5194/essd-12-197-2020, 2020.

Mao, K., Yuan, Z., Zuo, Z., Xu, T., Shen, X., and Gao, C.: Changes in Global Cloud Cover Based on Remote Sensing Data from 2003 to 2012, Chin. Geogr. Sci., 29, 306–315, https://doi.org/10.1007/s11769-019-1030-6, 2019.

Matsuura, K. and Willmott, C. J.: Terrestrial precipitation: 1900-2010 gridded monthly time series, University of Delaware, 2012.

Mohajane, M., Costache, R., Karimi, F., Bao Pham, Q., Essahlaoui, A., Nguyen, H., Laneve, G., and Oudija, F.: Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area, Ecological Indicators, 129, 107869, https://doi.org/10.1016/j.ecolind.2021.107869, 2021.

Peng, S., Ding, Y., Liu, W., and Li, Z.: 1 km monthly temperature and precipitation dataset for China from 1901 to 2017, Earth Syst. Sci. Data, 11, 1931–1946, https://doi.org/10.5194/essd-11-1931-2019, 2019.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle, Theor Appl Climatol, 115, 15–40, https://doi.org/10.1007/s00704-013-0860-x, 2014.

Serrano-Notivoli, R., Beguería, S., and de Luis, M.: STEAD: a high-resolution daily gridded temperature dataset for Spain, 18, 2019.

Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z., and Hong, Y.: Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets, Remote Sensing of Environment, 240, 111697, https://doi.org/10.1016/j.rse.2020.111697, 2020.

Xiao, C., Li, P., Feng, Z., and Wu, X.: Spatio-temporal differences in cloud cover of Landsat-8 OLI observations across China during 2013–2016, Journal of Geographical Sciences, 28, 429–444, https://doi.org/10.1007/s11442-018-1482-0, 2018.

Xu, Y., Knudby, A., Shen, Y., and Liu, Y.: Mapping Monthly Air Temperature in the Tibetan Plateau From MODIS Data Based on Machine Learning Methods, 11, 345–354, https://doi.org/10.1109/JSTARS.2017.2787191, 2018.

Yin, J., Guo, S., Gu, L., Zeng, Z., Liu, D., Chen, J., Shen, Y., and Xu, C.-Y.: Blending multi-satellite, atmospheric reanalysis and gauge precipitation products to facilitate hydrological modelling, Journal of Hydrology, 593, 125878, https://doi.org/10.1016/j.jhydrol.2020.125878, 2021.

Zhang, H., Zhang, F., Ye, M., Che, T., and Zhang, G.: Estimating daily air temperatures over the Tibetan Plateau by dynamically integrating MODIS LST data, 121, 11,425-11,441, https://doi.org/10.1002/2016JD025154, 2016.