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.

This study describes a new 1km dataset of monthly-mean, monthly-maximum and monthly-minimum surface temperature's over China, developed using machine learning methods. The method used for the final data set was chosen as the best performing method, after a comparison of three modern techniques. A dataset of 613 weather stations over China was used to train and test the machine learning methods. This study is very clearly written and the Figures are of high quality. I agree with all of reviewer 1's comments, so will not repeat these points and assume they have been addressed within the manuscript, but I will add a few further comments below.

Response: Many thanks for the constructive comments.

General comments:

Q1: There is always a tradeoff between spatial and temporal resolution when designing new data products. Can you explain in a few sentences in the manuscript why you chose to create a product with such high spatial resolution but such low temporal resolution? You make comparisons at the end to the ERA5 dataset which does have much lower spatial resolution (~30 x less) but it has hourly temporal resolution (720 x more) which is very useful for a number of applications. Comments suggesting the applications where you think this dataset may be preferable to the others mentioned would also be useful.

Response: Thanks a lot for your suggestion. Here are the reasons we produce the monthly product with high spatial resolution. First, the monthly temperature data is crucial for multiple studies and applications such as agriculture (Meshram et al., 2020), meteorological disasters (Tigkas et al., 2019) and ecology (Leihy et al., 2018). Second, the station data we obtained are from the China Meteorological Data Service Centre where the daily temperature data are not available. Thirdly, the daily temperature data with a high spatial resolution for a long period is enormously huge. Creating the data and storing the data for us is still quite challenging. ERA5 has high temporal resolution while the spatial resolution is low. We will mention this in the discussion session in the later revision of our paper.

Q2: You mention ERA5 is only available from 1979, but it is now available back to 1950, so could be used to incorporate dynamical variables (as suggested by reviewer 1). I'm not suggesting you do this, but in the limitations this could be a point for future development. And the text should be updated to reflect the availability of ERA5.

Response: The ERA5 is collected and processed on the Google Earth Engine platform, where the data is only available from 1979. We will update the year in the manuscript. Incorporating the ERA5 can be a good point for our future study. Thanks a lot for your advice. We will add some text to discuss ERA5 in the Discussion section.

Q3: Is any quality control performed on the meteorological station data you use as inputs? A few stations with low quality data could skew the results in data sparse regions.

Response: Yes. We train the model for each month. In each month, we deleted the stations with 999999 or 999998 values which mean the station in that month has no data. We also checked the value range of the air temperature in different months and all the selected stations are with reasonable values.

Q4: Do you know if the final model output is sensitive to the choice of stations used in the test/training dataset? I imagine that this could heavily influence the results in the data sparse regions.

Response: In order to find out if the model result is sensitive to the selection of weather stations used in the training and testing dataset, we conducted some experiments by randomly splitting the data into training and testing sets 50 times. We used the data from 1990, 2000, 2010 to do the case study. As shown in Figure 1, the RMSE varies slightly from different scenarios of the test/training dataset, while there is no obvious variation in R^2 (Figure 2). In our study, we split the stations into testing and training stations in ArcGIS, which has considered the spatial distribution of the weather stations. We will put the relationship between the choice of stations with the model output in the Discussion session.

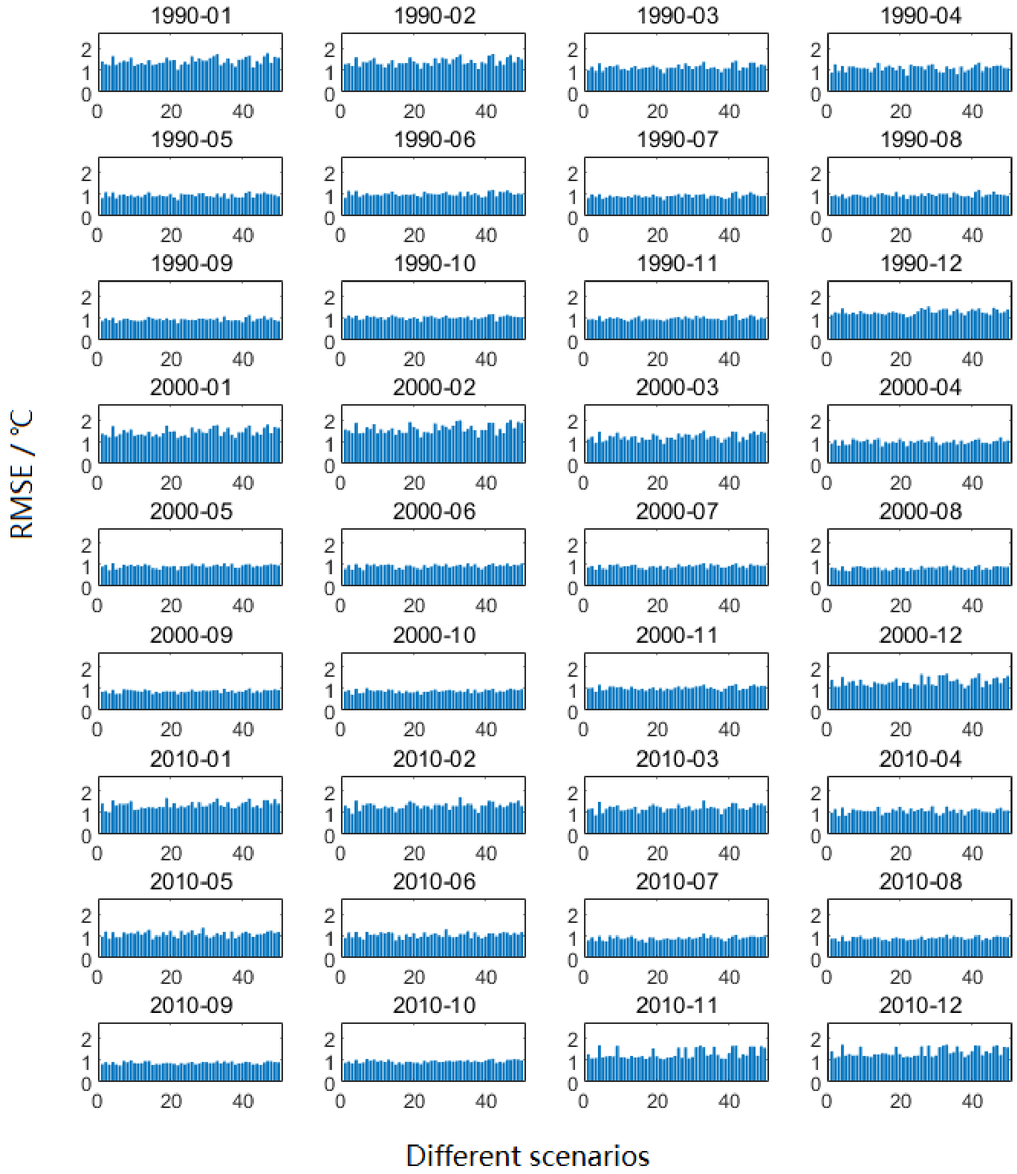

Figure 1 The RMSE using different testing and training datasets

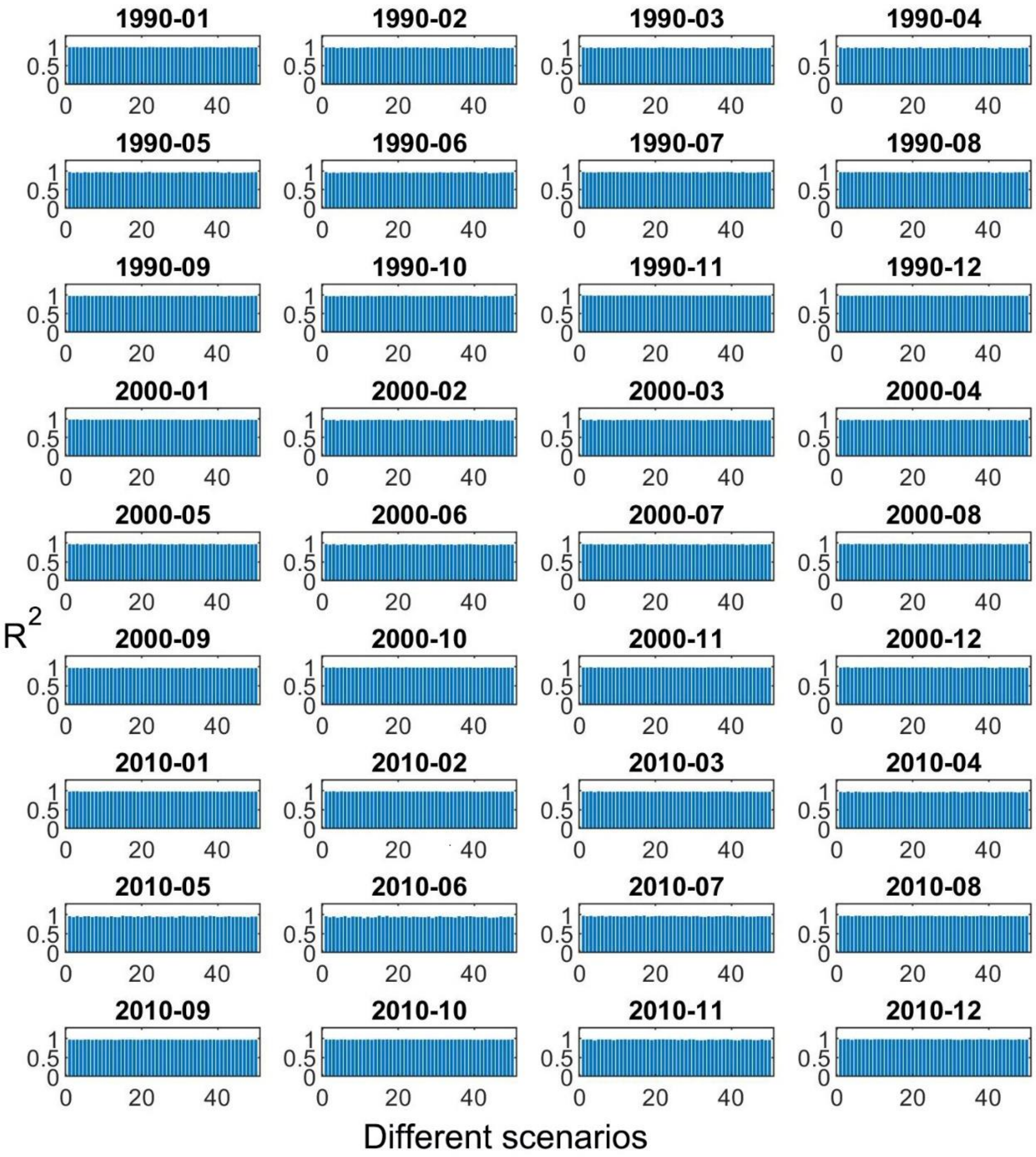

Figure 2 The R^2 using different testing and training datasets

Q5: Although you've included elevation, latitude and longitude there are multiple climatic regions in China, and a large amount of external drivers to variations in temperatures. The strength of these may modulate surface temperature behavior (e.g. the strength/location of the monsoon criculation, El Nino southern Oscillation, and other global teleconnections). Distance from the ocean could also play a role. Have you considered these in your explanations for months/stations with particuarly large residuals, or stations with strange behaviors? It could be that if a month had anomalous large scale weather conditions, which your machine learning methods are not trained to capture there are large residuals? These could make interesting case studies and could motivate future work incorporating some dynamical predictors.

Response: Many thanks for your comments. The global teleconnections can influence the surface temperature. This is a good topic for further study. We may use the data we generated combined with global teleconnections to do some research. In order to find out if the generated data in our study can capture the anomalous event, we did a case study in the Sichuan province. In 2006, there is an extremely severe drought event with an extremely high temperature in Sichuan (Li et al., 2011c). We extracted our maximum temperature data to the stations which were not used in the model training. We used nine stations in Sichuan and compared the mean temperature in July from 2003 to 2009. As shown in Figure 3, the temperature in 2006 is markedly higher than the neighbouring years, which means that our data can capture the anomalous condition.

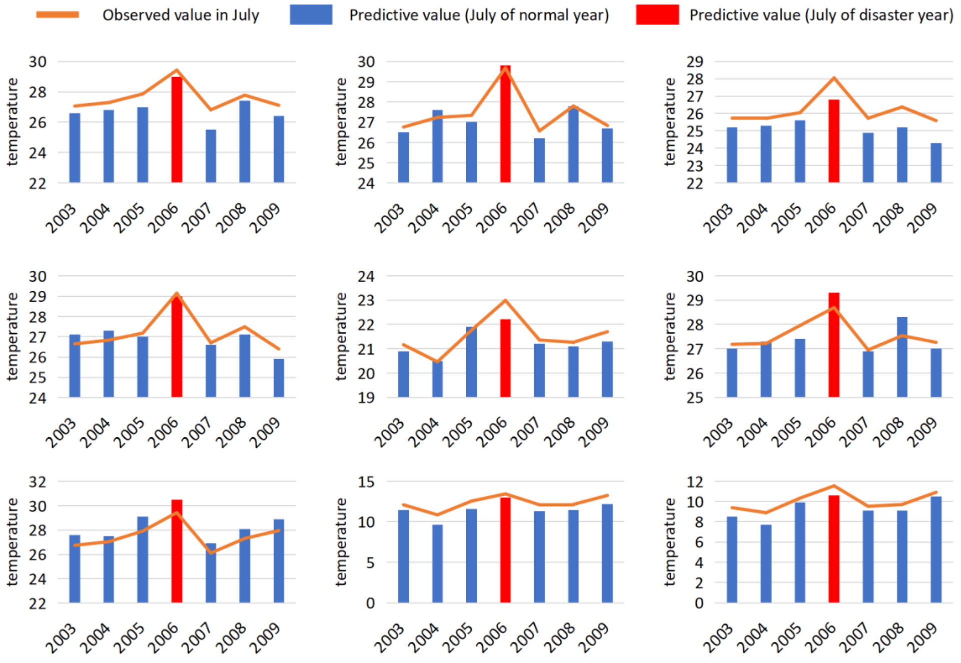

Figure 3 Maximum temperature for 9 different testing stations in Sichuan province

Q6: Figure 2: This is a nice depiction of the relationships. If you could briefly unpack the meteorological understanding behind this in the text it would be beneficial to readers. Have you checked that the relationships hold if different climatic regions of China are subset out?

Response: We used the subregions provided by Zhang et al. (Zhang et al., 2021). The subregions are divided according to the gradients of elevations and the precipitation patterns (Chen and Li, 2016; Zhang et al., 2021). The subregions are shown in Figure 4. We recalculated the correlation coefficients in each sub-region (Figure 5). The results show that Region 4, Region 5, Region 7 and Region 8 have clearly similar relationships as shown in Figure 2 in the manuscript. Since we train the model for the whole area of mainland China, the correlation in different sub-regions would not influence much about the whole region. It is an interesting topic to discuss different sub-regions, which can be helpful for regional studies. We would like to talk about this in the discussion session in our revision.

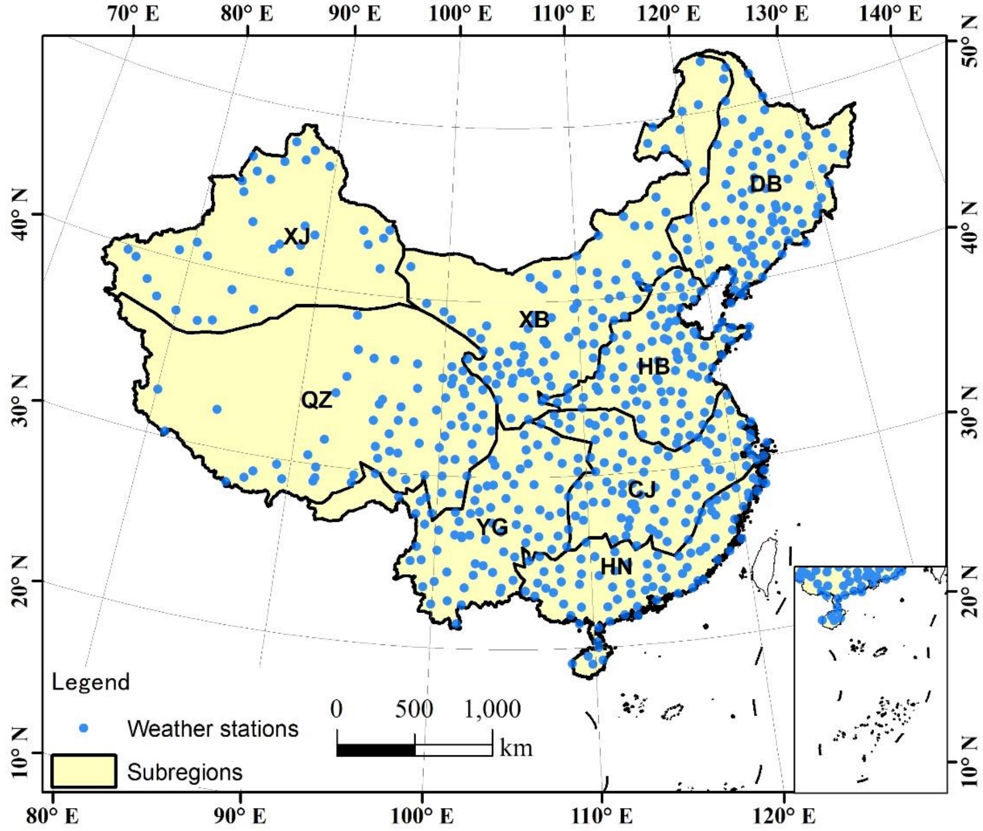

Figure 4 The division of the subregions, and the spatial distribution of the weather stations over the Chinese mainland.

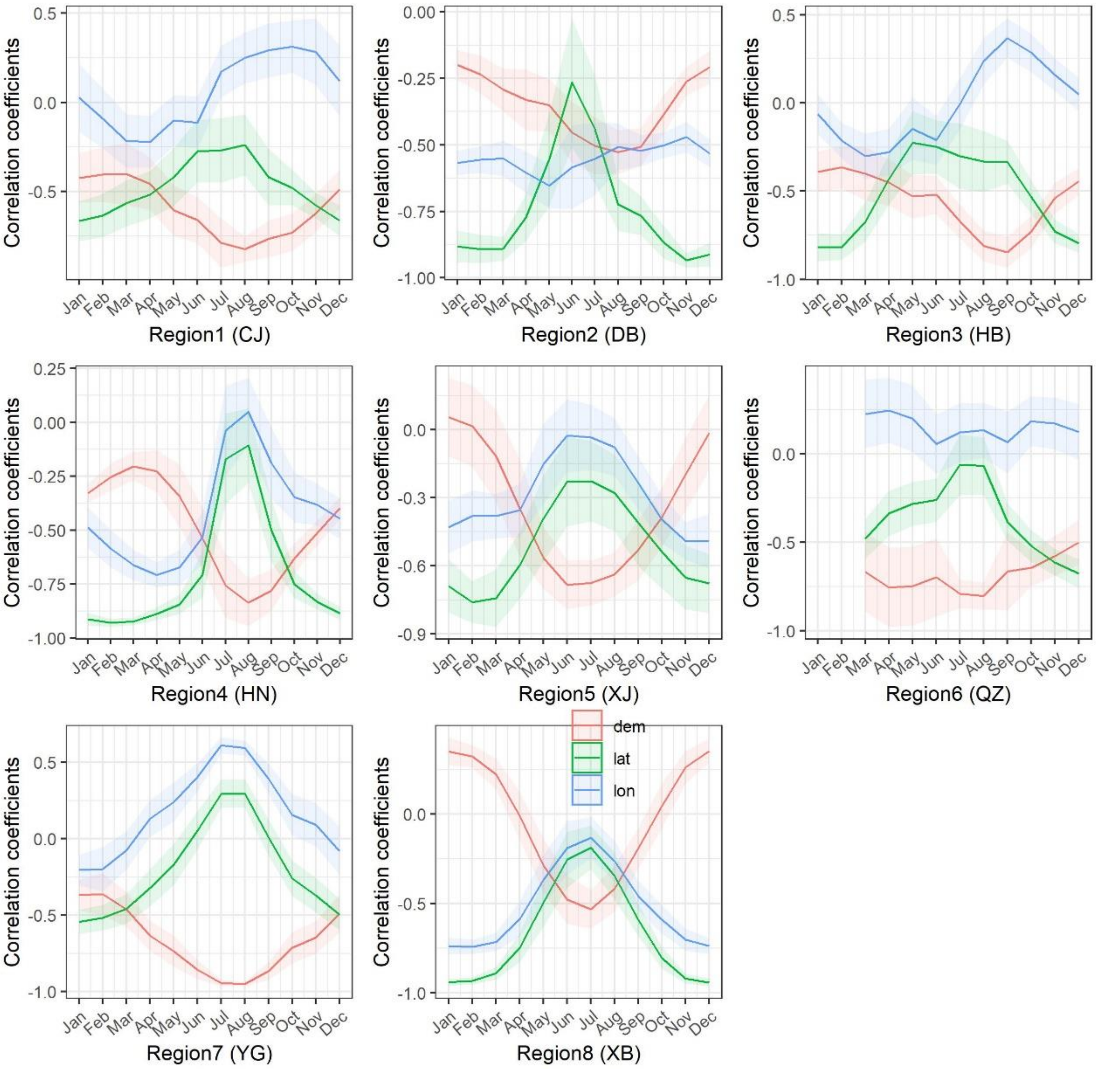

Figure 5 Correlation coefficients in different subregions

Q7: Line 190-195. So you have 840 different models. Can you comment on how different are all the 70 models for each month? (e.g. do all the January models look very similar?) This could be useful to understand if there are dynamical meteorological explanations for any outliers.

Response: We use the GPR model for all the 840 models for each month. The explicit basis in the GPR model is “constant” and the kernel function of the GPR algorithm is the exponential kernel. The predictor variables were standardized in the GPR. For each month, we use the temperature data of this month to train the model and then use this model to generate the grid data.

As the answer to Question 5: In order to find out if the generated data in our study can capture the anomalous event, we did a case study in the Sichuan province. In 2006, there is an extremely severe drought event with extreme high temperature in Sichuan (Li et al., 2011c). We extracted our maximum temperature data to the stations which were not used in the model training. We used nine stations in Sichuan and compared the mean temperature in July from 2003 to 2009. the temperature in 2006 is markedly higher than the neighboring years (Figure 6), which means that our data can capture the anomalous condition and it can be useful to understand the dynamical meteorological explanations for outliers.

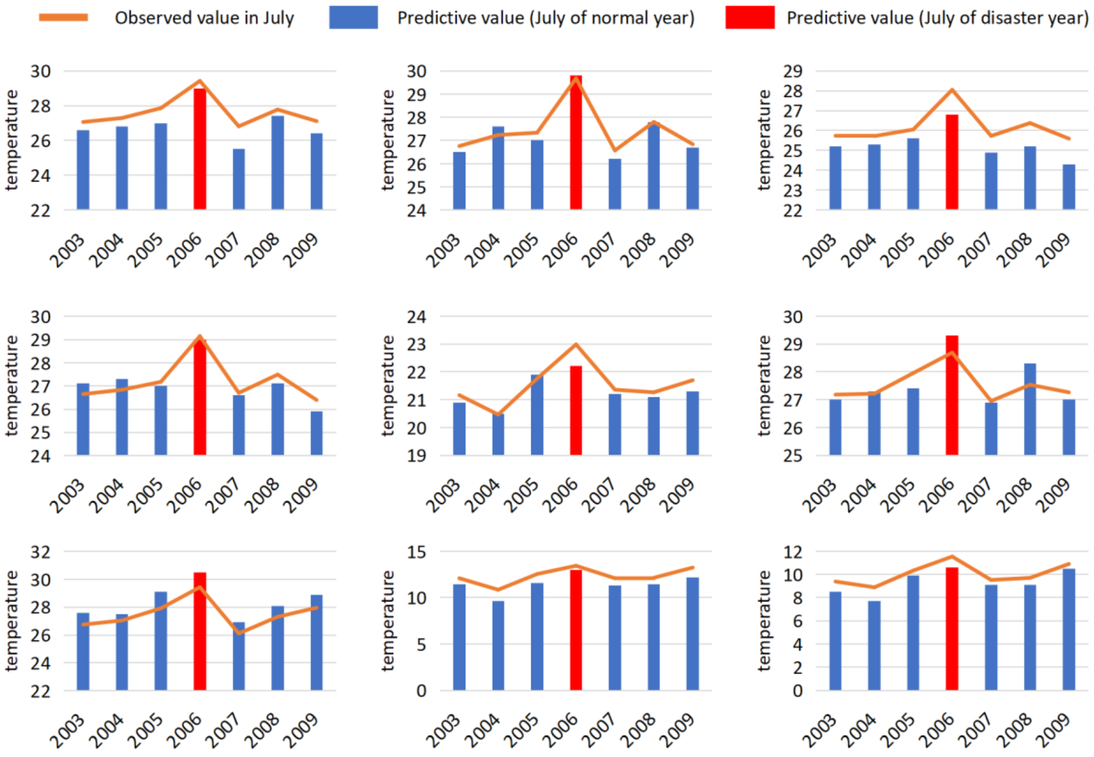

Figure 6 Maximum temperature for 9 different testing stations in Sichuan province

Q8: Line 243: Do you have a sense of why the errors are larger in the colder months? Are the impacts of local meteorological conditions larger in the cold season, which would make it more difficult for the methods to work? There may be meteorological literature on this.

Response: Thanks for your comments. As shown in Figure 2 in the manuscript, there are two variables (i.e., longitude and elevation) that have higher correlation with temperature, while in cold months there is only one variable with relatively higher correlation (i.e., latitude). We did some literature review. The large-scale mountain area is an important source of uncertainty in the temperature mapping, which can influence the spatial distribution of the surface air temperature such as in the area of the Qinghai Tibet Plateau and northwest China (Xu et al., 2018). The study of Stahl et al (Stahl et al., 2006) shows that the standard deviation for daily air temperature in winter is larger than that in summer. The study of Brunetti shows that the interpolation of high-resolution temperature for Italy has the lowest errors in spring and autumn and the highest errors in winter and explained that the elevation coefficients (lapse rates) are markedly different during winter (Brunetti et al., 2014). The air temperature in winter changes rapidly which may be a reason for the high estimation errors in winter (Amini et al., 2019). Rolland (Rolland, 2003) also found higher interpolation reliability for maximum and minimum temperature in winter than in summer. We will add the literature review in the Discussion session.

Q9: Figure 6: Can you comment on any features you're resolving here that are not seen in the lower resolution gridded products you will compare to? There are some very high resolution features on the map, but clarification that they are physical would be useful.

Response: The low spatial resolution can limit the ability to reflect the effects of complex topographies, land surface characteristics, and other processes on climate systems (Peng et al., 2019; Xu et al., 2017). The fine-scale data can provide realistic and reliable climate change information. We displayed the mean temperature of July, 2010 in the same region using ERA, FLADS datasets and the GPR dataset generated in our study (Figure 7). The GPR data can provide more spatial details than ERA and FLDAS (Figure 7).   We will add more clarification in our manuscript.

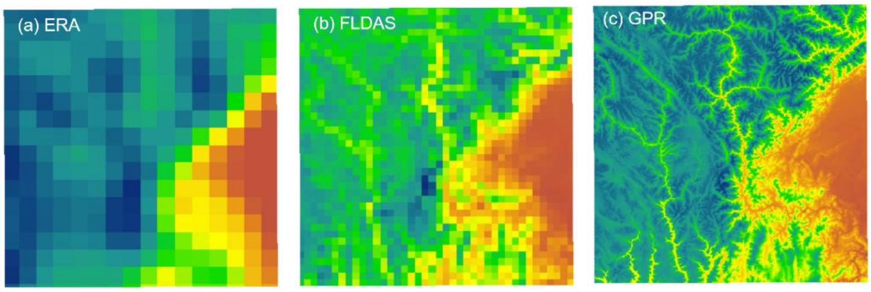

Figure 7 Comparison between the ERA, FLDAS and GPR datasets using the mean air temperature in July 2010

Q10: Does your trend analysis agree with the existing literature on global warming over China? If so include references to this.

Response: Thanks for your suggestion. We did a literature review and found some references which agree with the trend analysis in our study. The distribution of the temperature trend in China in our study agrees with the existing literature (Dong et al., 2015, p.1963–2012; Sun et al., 2018; You et al., 2021; Cui et al., 2017, p.1960–2015)

Q11: Figure 11: The Taylor diagrams show clear improvement from your new dataset. Also including some timeseries from locations not sampled from the observation network compared between the three datasets would be useful to understand how the four products sample the seasonal cycles of the variables.

Response: Many thanks for your advice. We randomly selected three locations not sampled from the observation network to make a comparison. The location of the points is shown in Figure 8. We extract the ERA, FLDAS and the GPR mean air temperature data to the three new points to make a comparison. As shown in Figure 9, the mean air temperature data of the three datasets have the similar pattern from January 1982 to December 2019. We do not have actual temperature data from the three randomly selected points, so it is not possible to make an accuracy comparison. However, the ERA and FLADS datasets are already used in a lot of studies (McNally et al., 2017; Hersbach et al., 2020), which means they are reliable to some extent. Thus the similar pattern between the GPR data and the ERA/FLDAS data can show the reliability of the GPR dataset we generated to a certain extent.

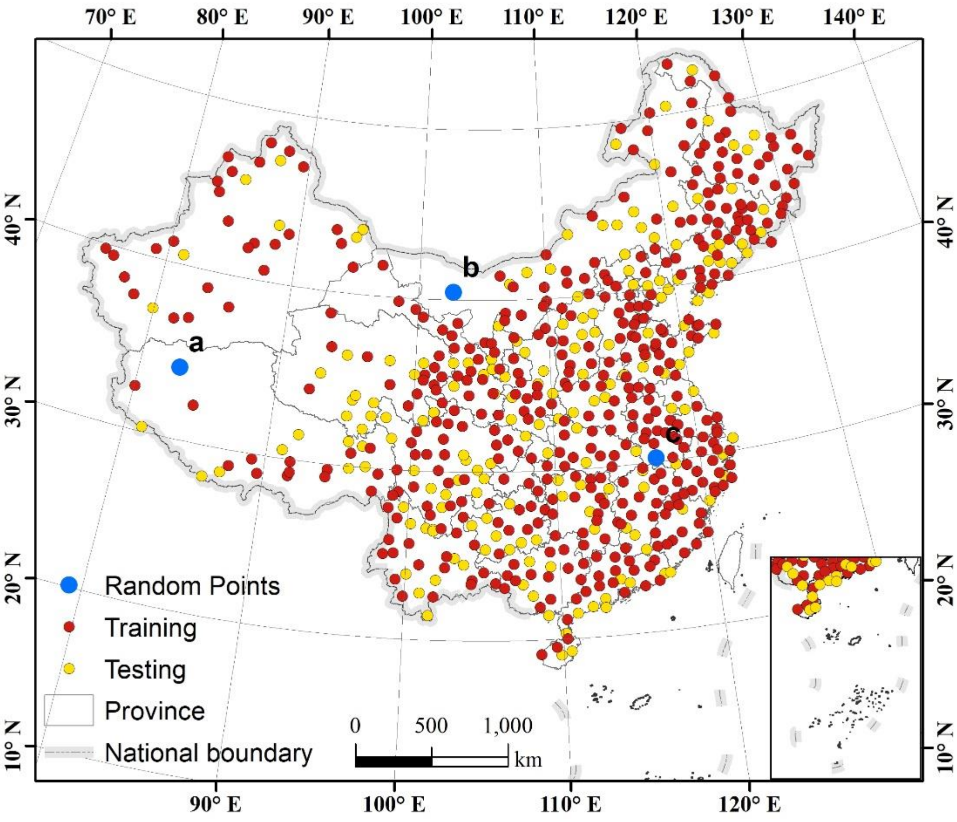

Figure 8 The location of the three random points

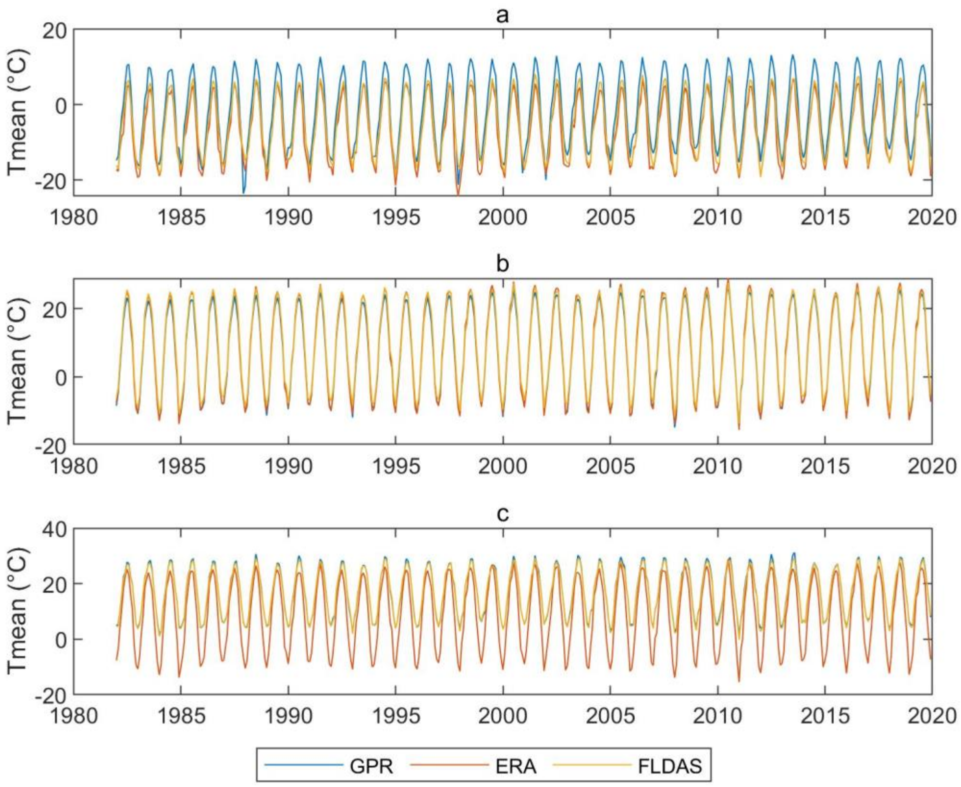

Figure 9 Comparisons of the data time series for mean air temperature in three random points from January 1982 to December 2019

Small corrections:

Q12: The height of the air temperatures (surface, 1.5m, 2m) should be added to the manuscript when this is mentioned.

Response: The height of the air temperatures from the weather stations is 2 m. We will add this to section 2.1 Meteorological station data.

Q13: The acronyms for datasets/methods should be defined in the abstract to make it easier to read.

Response: ANUSPLIN: (short for Australian National University Spline); TerraClimate: Monthly Climate and Climatic Water Balance for Global Terrestrial Surfaces; FLDAS: Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System, and ERA5: ECMWF Climate Reanalysis. We will add the full definition in the Abstract as you suggested.

Q14: Line 38: after commenting on the limitations of the observing stations you could comment here on the limitations of reanalysis based products.

Response: Thanks for the suggestion. The reanalysis based data usually have low spatial resolution, which limits their ability to reflect the effects of complex topographies, land surface characteristics, and other processes on climate systems (Peng et al., 2019; Xu et al., 2017). We will add this comment to our manuscript.

Q15: Throughout the text when you say ‘high resolution’ this should be changed to ‘high spatial resolution’ e.g. line 55.

Response: Yes. It should be changed to “high spatial resolution” to be clearer.

Q16: Line 56: ‘traditional interpolation techniques’ might be clearer?

Response: We agree with this. We will revise it in the manuscript.

Q17: Line 58-60: You comment on a few studies which talk about the superior performance of machine learning techniques but you do not say what the benchmark is that they’ve succeeded against. This should be included.

Response: The machine learning methods were selected mainly according to the applications of machine learning methods in previous studies. There is potential in applying machine learning methods to predict spatially continuous variables. The combination of machine learning and the traditional model can usually have better performance (Appelhans et al., 2015; Li et al., 2011b). Secondary information considered such as slope, latitude and longitude can improve the performance of machine learning as they provide essential information for machine learning methods. (Li et al., 2011b; Alizamir et al., 2020; Appelhans et al., 2015; Kisi et al., 2017; Zhu et al., 2018). We will include more descriptions in the revised manuscript.

Q18: Line 61: 'estimation of short-term air temperature ' – I'm not sure what you mean by this?

Response: We are sorry for making you confused. Here we should delete "short-term" to make it clearer.

Q19: Line 86: The link here gives me an Error 404.

Response: The link is not available recently, because the data were not available online. We will change our link to the homepage (https://data.cma.cn/data/).

Q20: Around the discussion for Figure 1 it would be interesting to know the spatial distance between observation sites. This might be a small indication of confidence in the final machine learning model output.

Response: Thanks for your advice. We generated the Euclidean distance for the observation stations. Figure 10 shows that the Euclidean distance is quite small in most of the region. The Euclidean distance is relatively large in the west of the Tibetan Plateau and the small region in Inner Mongolia. The larger Euclidean distance means the stations in that region are sparse, which can have an impact on the interpolation accuracy (Hijmans et al., 2005; Li et al., 2011a). We will talk about the model output and the spatial distance in the revised manuscript.

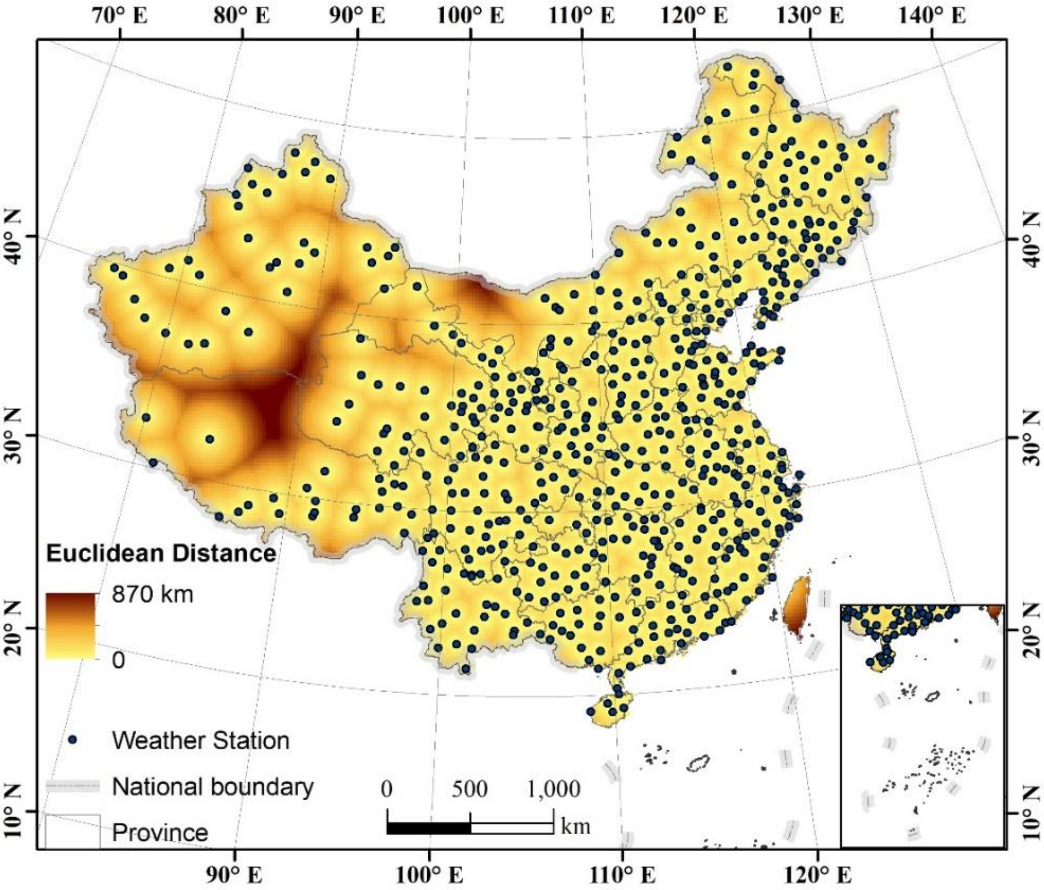

Figure 10 Euclidean distance of the weather stations

Q21: Section 2.3: When commenting on the spatial resolution of the gridded products used for comparison it would be useful to also have this in km.

Response: Thanks for your suggestion. The spatial resolution of TerraClimate, FLDAS, and ERA5 are about 4.6 km, 11 km and 28 km, respectively. We will update the text in the manuscript according to your suggestion.

Q22: Line 149: ‘machining learning’ should be ‘machine learning’

Response: We will change the wrong spelling.

Q23: Section 3.2.2 Are the choices of parameters for the SVM method standard in the literature? Can you please comment on your choices?

Response: Thanks for your suggestion. For the SVM models, we used the Gaussian kernel. The kernel scale is set to 1.7 for all SVM models. Hyperparameters are important for the performance of the model. We optimized the kernel scale of SVM and made a comparison. The box constraint and the epsilon hyperparameters are varying from month to month according to the training data of each month are different. The mean temperature data from January to December from 1951 to 2020 was used for comparing the new model with varying kernel scales and the used model in our study. The accuracy of each month between the optimized model and the used model in the paper is similar (Figure 11), the optimized models do not improve significantly. As we can see that the adjustment of the hyperparameters has little impact on the model accuracy in our study. Besides, the models used in our study are more time-saving and efficient.

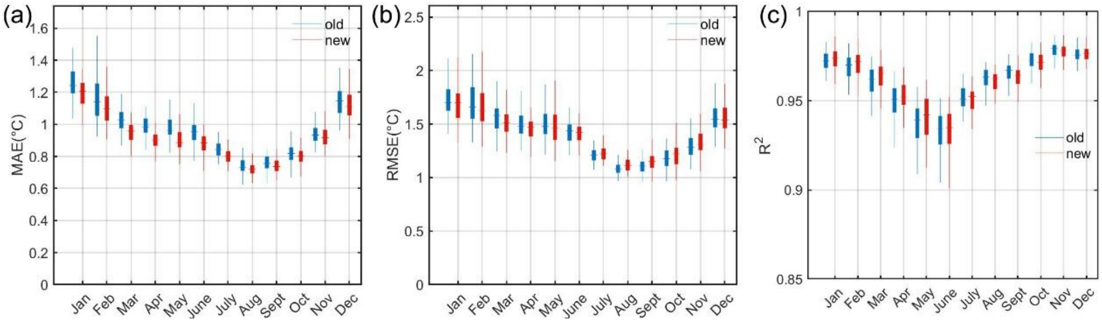

Figure 11 Comparison between optimized SVM model (new) and the used model in the paper (old).

Q24: Line 342: ‘ shows a cyclic pattern’ might be clearer.

Response: We will change the text as suggested.

References:

Alizamir, M., Kisi, O., Ahmed, A. N., Mert, C., Fai, C. M., Kim, S., Kim, N. W., and El-Shafie, A.: Advanced machine learning model for better prediction accuracy of soil temperature at different depths, PLOS ONE, 15, e0231055, https://doi.org/10.1371/journal.pone.0231055, 2020.

Amini, M. A., Torkan, G., Eslamian, S., Zareian, M. J., and Adamowski, J. F.: Analysis of deterministic and geostatistical interpolation techniques for mapping meteorological variables at large watershed scales, Acta Geophys., 67, 191–203, https://doi.org/10.1007/s11600-018-0226-y, 2019.

Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., and Nauss, T.: Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania, Spatial Statistics, 14, 91–113, https://doi.org/10.1016/j.spasta.2015.05.008, 2015.

Brunetti, M., Maugeri, M., Nanni, T., Simolo, C., and Spinoni, J.: High-resolution temperature climatology for Italy: interpolation method intercomparison, 34, 1278–1296, https://doi.org/10.1002/joc.3764, 2014.

Chen, F. and Li, X.: Evaluation of IMERG and TRMM 3B43 Monthly Precipitation Products over Mainland China, 8, 472, https://doi.org/10.3390/rs8060472, 2016.

Cui, L., Wang, L., Lai, Z., Tian, Q., Liu, W., and Li, J.: Innovative trend analysis of annual and seasonal air temperature and rainfall in the Yangtze River Basin, China during 1960–2015, Journal of Atmospheric and Solar-Terrestrial Physics, 164, 48–59, https://doi.org/10.1016/j.jastp.2017.08.001, 2017.

Dong, D., Huang, G., Qu, X., Tao, W., and Fan, G.: Temperature trend–altitude relationship in China during 1963–2012, Theor Appl Climatol, 122, 285–294, https://doi.org/10.1007/s00704-014-1286-9, 2015.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

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.

Kisi, O., Sanikhani, H., and Cobaner, M.: Soil temperature modeling at different depths using neuro-fuzzy, neural network, and genetic programming techniques, Theor Appl Climatol, 129, 833–848, https://doi.org/10.1007/s00704-016-1810-1, 2017.

Leihy, R. I., Duffy, G. A., Nortje, E., and Chown, S. L.: High resolution temperature data for ecological research and management on the Southern Ocean Islands, Sci Data, 5, 180177, https://doi.org/10.1038/sdata.2018.177, 2018.

Li, J., Heap, A. D., Potter, A., and Daniell, J. J.: Application of machine learning methods to spatial interpolation of environmental variables, Environmental Modelling & Software, 26, 1647–1659, https://doi.org/10.1016/j.envsoft.2011.07.004, 2011a.

Li, J., Heap, A. D., Potter, A., Huang, Z., and Daniell, J. J.: Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin, Continental Shelf Research, 31, 1365–1376, https://doi.org/10.1016/j.csr.2011.05.015, 2011b.

Li, Y., Xu, H., and Liu, D.: Features of the extremely severe drought in the east of Southwest China and anomalies of atmospheric circulation in summer 2006, Acta Meteorol Sin, 25, 176–187, https://doi.org/10.1007/s13351-011-0025-8, 2011c.

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., Funk, C., Peters-Lidard, C. D., and Verdin, J. P.: A land data assimilation system for sub-Saharan Africa food and water security applications, Sci Data, 4, 170012, https://doi.org/10.1038/sdata.2017.12, 2017.

Meshram, S. G., Kahya, E., Meshram, C., Ghorbani, M. A., Ambade, B., and Mirabbasi, R.: Long-term temperature trend analysis associated with agriculture crops, Theor Appl Climatol, 140, 1139–1159, https://doi.org/10.1007/s00704-020-03137-z, 2020.

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.

Rolland, C.: Spatial and Seasonal Variations of Air Temperature Lapse Rates in Alpine Regions, 16, 1032–1046, https://doi.org/10.1175/1520-0442(2003)016<1032:SASVOA>2.0.CO;2, 2003.

Stahl, K., Moore, R. D., Floyer, J. A., Asplin, M. G., and McKendry, I. G.: Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density, Agricultural and Forest Meteorology, 139, 224–236, https://doi.org/10.1016/j.agrformet.2006.07.004, 2006.

Sun, X., Ren, G., Ren, Y., Fang, Y., Liu, Y., Xue, X., and Zhang, P.: A remarkable climate warming hiatus over Northeast China since 1998, Theor Appl Climatol, 133, 579–594, https://doi.org/10.1007/s00704-017-2205-7, 2018.

Tigkas, D., Vangelis, H., and Tsakiris, G.: Drought characterisation based on an agricultureoriented standardised precipitation index, Theor Appl Climatol, 135, 1435–1447, https://doi.org/10.1007/s00704-018-2451-3, 2019.

Xu, C., Wang, J., and Li, Q.: A New Method for Temperature Spatial Interpolation Based on Sparse Historical Stations, 31, 1757–1770, https://doi.org/10.1175/JCLI-D-17-0150.1, 2018.

Xu, J., Gao, Y., Chen, D., Xiao, L., and Ou, T.: Evaluation of global climate models for downscaling applications centred over the Tibetan Plateau, 37, 657–671, https://doi.org/10.1002/joc.4731, 2017.

You, Q., Cai, Z., Wu, F., Jiang, Z., Pepin, N., and Shen, S. S. P.: Temperature dataset of CMIP6 models over China: evaluation, trend and uncertainty, Clim Dyn, 57, 17–35, https://doi.org/10.1007/s00382-021-05691-2, 2021.

Zhang, L., Li, X., Zheng, D., Zhang, K., Ma, Q., Zhao, Y., and Ge, Y.: Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach, Journal of Hydrology, 594, 125969, https://doi.org/10.1016/j.jhydrol.2021.125969, 2021.

Zhu, S., Nyarko, E. K., and Hadzima-Nyarko, M.: Modelling daily water temperature from air temperature for the Missouri River, PeerJ, 6, e4894, https://doi.org/10.7717/peerj.4894, 2018.