Response to Referee #2 for the Second review

We sincerely appreciate your second review and the insightful comments regarding our manuscript entitled "Permafrost temperature baseline at 15 meters depth in the Qinghai-Tibet Plateau (2010–2019)" (MS No.: essd-2024-114). Your feedback has been invaluable in improving the quality of the manuscript. We have thoroughly addressed all the points raised and have made the necessary revisions accordingly.

General comments

I would like to acknowledge the effort the authors have put into improving the manuscript according to both reviewers' comments. They have clearly taken the time to consider concerns and redo analysis to strengthen their study.

Unfortunately, I am concerned about the decisions made in terms of the appropriate machine learning algorithm. I appreciate that the authors have added the results from the other viable machine learning models they mention in the original manuscript, but I struggle with the decision and the arguments as to why the much better performing random forest model was dismissed. As I am not an expert in machine learning myself, I do not feel qualified to provide in-depth feedback about machine learning algorithms and their performance metrics, but if the statistics used to evaluate the model are R2, bias, and RMSE, I cannot understand how the SVR can be chosen as the best model when it's clearly performing poorly compared to the random forest algorithm.

While I understand the qualitative reasoning and I believe that the authors know the area better than me, I do not think that this alone justifies the choice of model. Are there any other statistics or performance metrics that you could use to evaluate the model and back up your qualitative results numerically? If you say you do not trust the RF model because of the reasons given in your response letter, how can you be sure that a much worse performing algorithm gives you better results?

I further wonder if the results of testing the other machine learning algorithms should be included in the paper as it is crucial for full transparency and reproducibility. Again,

not being an expert in machine learning, I am unsure about the best way to do so, but in its current state, the decisions made by the authors are not fully transparent and comprehensible.

I am happy to take another look at the manuscript after this pressing issue has been addressed. To do so, it would also be very helpful to have a full track changes version of the manuscript, including deleted/changed text between the two manuscripts. Right now I can only see additions and they come in two different colors, which is confusing.

I am aware that my comments may mean a lot of extra work for the authors, but I think it is important for the transparency and reproducibility of your results and its translation into future work and different regions in the permafrost area.

Response:

In predicting $MAGT_{15m}$ in this study, the Random Forest (RF) method highlights a common issue where the model's performance may not align with the geographic significance of the results. Despite RF shows a high R^2, its predicted values deviate from the actual physical significance due to the narrow distribution range of $MAGT_{15m}$ values, with a minimum of only -3.2 °C and a maximum of 0.9 °C. This limitation arises from RF's over-reliance on the range of the training data and its lack of extrapolation capability. As an ensemble algorithm based on decision trees, RF is adept at capturing complex nonlinear relationships within the data. However, its reliance on splitting rules may result in insufficient predictive power when faced with extreme values or boundary conditions.

RF tends to be highly dependent on the specific distribution of the training data, which makes it susceptible to being influenced by local data structures while neglecting broader geographic trends in ground temperature. This may explain the narrow range in the $MAGT_{15m}$ predictions. In other words, the performance of RF is largely depended on the representativeness of the training samples.

Figure R1 compares the elevation distribution of borehole locations used in the

manuscript with that of the Qinghai-Tibet Plateau (QTP) permafrost regions. Specifically, the comparison reveals that $MAGT_{15m}$ samples are primarily concentrated in areas below 5200 m. Due to challenging environmental conditions and limited access to high-altitude regions, sample points are sparse at elevations above this threshold. This makes it crucial to rely on models that can extrapolate $MAGT_{15m}$ values to these higher regions. However, the RF method demonstrates limited capability in this regard.

A comparison shows that the RF-predicted $MAGT_{15m}$ range (-3.2 to 0.9 °C) closely aligns with the observed range (-4.0 to 1.5 °C). This reflects a characteristic of the RF method: while it excels at modeling local patterns, it is strongly influenced by the available observations and fails to capture variations in extremely low ground temperatures. Based on the observed lapse rate of $MAGT_{15m}$ (0.4-0.9°C/100m) in the QTP permafrost regions, the $MAGT_{15m}$ at 6000 m is conservatively estimated to range from -6 °C to -10 °C. Since regions above 5200 m account for 22.0 % of the QTP permafrost areas, accurate predictions for these higher-altitude regions are crucial for a comprehensive analysis.

The support vector regression (SVR) method, on the other hand, predicts an extremely low value of -12.2 °C, which aligns well with the lapse rate-based estimate. The strength of SVR lies in its ability to capture global trends by identifying an optimal hyperplane in a high-dimensional space. SVR tends to model overarching trends in the data rather than focusing on local variations. While its R^2 value is somewhat lower, SVR effectively captures the spatial trends, which is a key advantage for ground temperature predictions in the QTP, especially when accounting for temperature gradients and spatial variations at high altitudes.

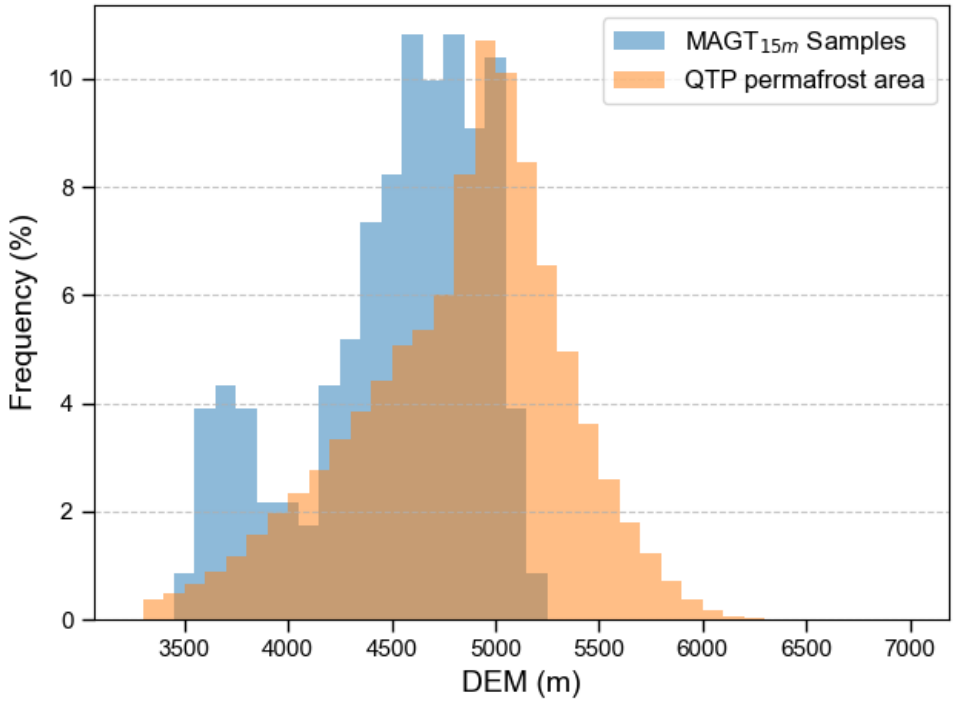

Figure R1: *Percentage frequency distribution of elevations for $MAGT_{15m}$ borehole locations compared to the permafrost regions of the Qinghai-Tibet Plateau.*

To further explore how different parameters in the RF method influence prediction results, we designed a series of sensitivity experiments. Three key parameters were identified in the RF method: ***mtry***, ***ntree***, and ***maxnodes***.

The ***mtry*** parameter defines the number of predictor variables randomly sampled to be considered for each split in a decision tree. It influences the diversity of the individual trees in the forest. A smaller value increases the model's variance, as it allows trees to explore different splits based on a smaller set of features. Conversely, a larger value leads to more similar trees, reducing variance but potentially increasing overfitting. For regression problems, it is typically set to one-third of the total features. In this study, *mtry* was set to 3, given that there are 10 predictor variables in total.

The ***ntree*** parameter specifies the number of decision trees to grow in the RF, essentially controlling the size of the forest. A larger value generally improves model stability and prediction accuracy by averaging over a greater number of trees. However, this also increases computation time and memory usage. The trade-off lies between

model accuracy and computational efficiency. Typical values often range from 500 to 1000, as seen in previous studies (Ran et al., 2021, 2022). To assess the impact of *ntree* on model performance, we varied it across a range of values: 20, 50, 200, 500, 1000, 2000, and 5000, evaluating its effects on accuracies and the spatial patterns of the predicted results.

The ***maxnodes*** parameter limits the maximum number of terminal nodes (leaves) that each individual tree in the forest can have, controlling the depth and complexity of the trees. A smaller value results in simpler trees with fewer splits, which can reduce overfitting and improve the model's generalization ability, but may introduce bias. Larger values allow trees to grow deeper and capture more complex patterns, but can lead to overfitting, especially with smaller datasets. The optimal value depends on the data and should be fine-tuned through cross-validation. Typically, values between 20 and 50 are chosen. To evaluate the RF model's performance under extreme conditions, we manually set *maxnodes* to 10, 20, 30, 40, 50, 100, and 200.

Table R1 summarizes the R^2 values of the RF model for 49 different combinations of *ntree* and *maxnodes* parameters. Each R^2 was evaluated using a 9:1 data split, with results averaged over 200 runs to ensure robustness. The analysis shows that parameter selection significantly impacts model performance, with R^2 values ranging from 0.66 to 0.92. With *ntree* held constant, increasing *maxnodes* (from 10 to 200) notably improved R^2 values. In contrast, when *maxnodes* was held constant, increasing *ntree* (from 20 to 5000) resulted in only a slight improvement in R^2. For RMSE, the values across the 49 parameter combinations ranged from 0.31 to 0.60 (Table R2). Increasing *maxnodes* notably reduced RMSE values, while increasing *ntree* led to only a slight reduction in RMSE.

Table R1: *R^2 statistics of the Random Forest (RF) model for various combinations of ntree and maxnodes parameters.*

ntree	maxnodes						
	10	20	30	40	50	100	200

20	0.66	0.76	0.82	0.86	0.89	0.90	0.90
50	0.69	0.78	0.84	0.87	0.89	0.92	0.91
200	0.70	0.79	0.84	0.88	0.90	0.92	0.92
500	0.69	0.78	0.84	0.87	0.90	0.92	0.92
1000	0.69	0.78	0.84	0.88	0.90	0.92	0.92
2000	0.68	0.80	0.84	0.88	0.90	0.92	0.92
5000	0.69	0.78	0.84	0.88	0.90	0.92	0.92

Table R2: *RMSE statistics of the Random Forest (RF) model for various combinations of ntree and maxnodes parameters.*

ntree	maxnodes						
	10	20	30	40	50	100	200
20	0.60	0.50	0.44	0.41	0.36	0.35	0.34
50	0.60	0.50	0.43	0.39	0.36	0.31	0.33
200	0.58	0.49	0.43	0.37	0.35	0.31	0.32
500	0.59	0.50	0.43	0.39	0.35	0.32	0.31
1000	0.58	0.50	0.43	0.38	0.36	0.32	0.32
2000	0.59	0.49	0.43	0.39	0.35	0.31	0.31
5000	0.59	0.49	0.42	0.38	0.36	0.32	0.32

Variations in parameter combinations lead to differences in evaluation metrics. Nonetheless, even the lowest R^2 (0.66) and highest RMSE (0.60 °C) achieved by the RF method outperform those of SVR method (R^2=0.48, RMSE=0.71). From an accuracy evaluation perspective, the RF method demonstrates robust reliability and offers further potential for improvement through parameter optimization. However, since the RF method is constrained by the distribution of sample points and fails to accurately predict lower $MAGT_{15m}$ values, we are also interested in exploring whether different parameter choices could help address this limitation.

To this end, we analyzed the impact of 49 parameter combinations on the maximum (Table R3) and minimum (Table R4) $MAGT_{15m}$ predictions for the QTP permafrost regions. The results reveal that parameter variations have minimal impact on the extreme values, with maximum predictions consistently ranging from 0.1 °C and 1.1 °C and minimum predictions spanning -2.5 °C to -3.4 °C. These findings suggest that even with optimized parameter settings, the RF method fails to generate reasonable spatial predictions based on the existing sample data.

Table R3: *Maximum value statistics of the Random Forest (RF) model for various combinations of ntree and maxnodes parameters.*

ntree	maxnodes						
	10	20	30	40	50	100	200
20	0.32	0.33	0.75	1.11	0.95	0.96	0.87
50	0.30	0.63	0.87	0.84	0.82	0.84	0.98
200	0.10	0.56	0.75	0.82	0.91	0.82	0.86
500	0.17	0.55	0.71	0.75	0.86	0.85	0.87
1000	0.17	0.54	0.73	0.81	0.85	0.84	0.84
2000	0.16	0.55	0.71	0.80	0.85	0.87	0.86
5000	0.15	0.53	0.71	0.79	0.82	0.85	0.85

Table R4: *Minimum value statistics of the Random Forest (RF) model for various combinations of ntree and maxnodes parameters.*

ntree	maxnodes						
	10	20	30	40	50	100	200
20	-2.75	-3.10	-3.32	-3.16	-3.42	-3.26	-3.33
50	-2.75	-2.91	-2.98	-3.04	-3.22	-3.01	-3.08
200	-2.51	-2.91	-2.99	-3.04	-3.07	-3.12	-3.01
500	-2.60	-2.86	-3.03	-3.14	-3.06	-3.12	-3.12

1000	-2.59	-2.85	-3.09	-3.09	-3.06	-3.09	-3.16
2000	-2.59	-2.86	-2.98	-3.08	-3.10	-3.11	-3.14
5000	-2.60	-2.88	-3.03	-3.05	-3.10	-3.13	-3.12

The RF method is not unique in its ability to capture local patterns. Further testing revealed that other commonly used machine learning algorithms, such as eXtreme Gradient Boosting (XGBoost) and K-Nearest Neighbors (KNN), exhibit similar behavior to the RF method (Figure R2). Under the same data samples, the minimum and maximum values predicted by XGBoost and KNN methods range from -3.3 to -4.0 °C and from 0.9 to 1.5 °C, respectively.

Like RF, both XGBoost and KNN rely on the distribution of the training data for modeling, and their predictions are primarily shaped by the patterns present within the existing data. Whether through the ensemble of decision trees in RF, the gradient boosting mechanism in XGBoost, or the neighbor-based approach in KNN, these models struggle to predict values beyond the sample range. The prediction results of these algorithms generally do not extend outside the distribution range of the training data, particularly when predicting extreme values (either minimum or maximum). These models tend to replicate the boundaries of the training data, making it difficult to accurately forecast values outside of this range. The primary objective of these models is to fit the training data, rather than extrapolate. While they optimize performance by minimizing training error, they lack inherent mechanisms for extrapolation beyond the data distribution.

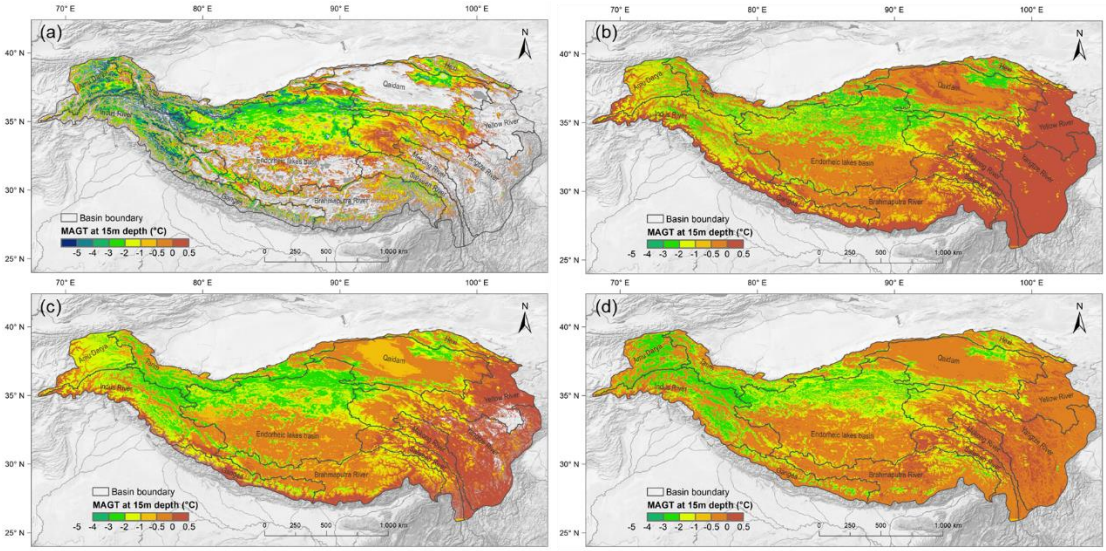

Figure R2: *Spatial distribution of predicted mean annual ground temperatures at the 15m depth ($MAGT_{15m}$) across the Qinghai-Tibet Plateau during 2010-2019, based on support vector regression (a), extreme gradient boosting (b), random forest (c) and K-nearest neighbors (d) models.*

In summary, the limited number of sample points in higher-altitude regions, where $MAGT_{15m}$ values are generally lower, prevents the RF method from accurately capturing spatial trends in these areas based solely on the existing dataset. Due to the harsh climatic conditions at high altitudes, obtaining observational data in these regions remains challenging with current observational capabilities. As a result, there is still a need to leverage models capable of extrapolation to estimate ground temperature trends in these high-altitude areas.

In response to the question, "***Are there any other statistics or performance metrics that you could use to evaluate the model and back up your qualitative results numerically?***" — addressing this issue poses a considerable challenge. Due to the limited availability of additional observational data and reliable spatial information, identifying alternative statistical metrics to evaluate the model is currently not feasible.

In the early stages of research on spatializing ground temperatures in high-altitude permafrost regions, researchers often used three-dimensional zonality based on longitude, latitude, and elevation for spatial prediction (Nan et al., 2013). While this approach may introduce systematic errors, particularly when predicting extreme high

and low values, it effectively captures the overall spatial trends of ground temperature on the QTP.

Here, we also employed multiple linear regression, incorporating longitude, latitude, and elevation as variables, to establish the spatial distribution of $MAGT_{15m}$. For ease of comparison, this prediction is referred to as LLE in the following discussion. The LLE result was used as the ground truth assumption to evaluate the performance of the RF method and other approaches.

Figure R3 illustrates the percentage frequency distribution of $MAGT_{15m}$ predicted by different methods. The comparison shows that the histograms of the support vector regression (SVR) and generalized additive model (GAM) methods closely resemble that of LLE method, exhibiting a high degree of consistency (Figure R3 a). In contrast, the histograms of the RF and XGBoost methods deviate significantly from LLE. The values predicted by RF and XGBoost are mainly concentrated between -3 °C and 1 °C, with notable peaks around 0 °C and 2 °C, where their frequency is substantially higher than in the LLE distribution. It is evident that the predictions from the SVR and GAM methods align well with the broader spatial temperature pattern of the QTP, as represented by LLE method. On the other hand, methods such as the RF and XGBoost fail to accurately capture this overarching spatial trend.

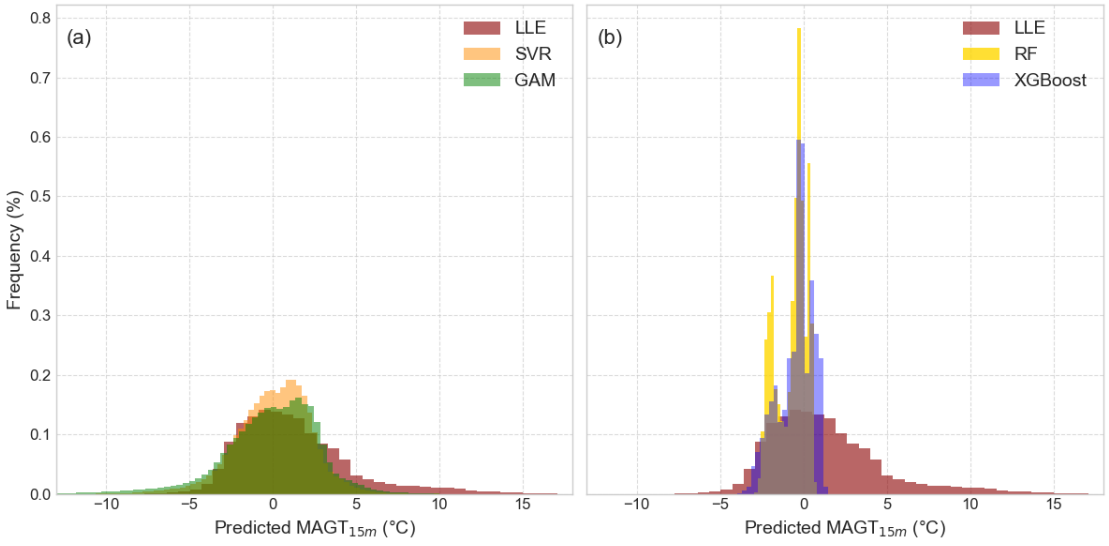

Figure R3: *Percentage frequency distribution of predicted $MAGT_{15m}$ across the Qinghai-Tibet*

Plateau permafrost regions with different methods (LLE-three dimensional zonality method based on longitude, latitude and elevation; SVR-support vector regression, GAM-generalized additive model, RF-random forest and XGBoost- extreme gradient boosting).

In addition to comparing the distribution patterns of the histograms, we used the LLE result as the ground truth to assess the accuracy of the four methods (Table R5). The SVR method exhibited the lowest errors, with ME, MAE, and RMSE values of 1.36, 1.60, and 2.60 °C, respectively, along with the highest R^2 value of 0.46. The GAM method showed slightly higher errors than SVR, resulting in a slightly lower R^2 value of 0.39. In contrast, the R^2 values for RF and XGBoost were only -0.05 and 0.03, respectively, accompanied by relatively larger errors. This suggests that the spatial predictions from these methods are only weakly related to the spatial variation trend observed in the LLE method.

Table R5: *Accuracy evaluation of the four models compared to the result of LLE method*[*]*.*

Performance	SVR	GAM	RF	XGBoost
R^2	0.46	0.39	-0.05	0.03
ME (°C)	1.36 (±2.21)	1.37 (±2.41)	1.99 (±3.04)	1.76 (±3.00)
MAE (°C)	1.60 (±2.04)	1.67 (±2.21)	2.32 (±2.79)	2.23 (±2.67)
RMSE (°C)	2.60 (±6.53)	2.77 (±8.77)	3.63 (±8.19)	3.47 (±8.07)

[*]*SVR, support vector regression; GAM, generalized additive model; RF, random forest; XGBoost, extreme gradient boosting.* R^2*, ME, MAE and RMSE with 1 standard deviation.*

Although the LLE method has been used as a reference for comparing the performance of different machine learning models, it is important to note that the evaluation remains largely qualitative. This is because the LLE result is also empirical in nature, commonly used in earlier studies with limited data availability. As such, it does not represent a strictly quantitative assessment of accuracy. Nevertheless, this comparison offers a valuable new perspective for evaluating the performance of different models. Ultimately, the comparisons conducted above clearly indicate that the

RF model, when applied to the current ground temperature dataset, is not optimally suited for predicting permafrost temperatures in the QTP.

The initial selection of the SVR method in this study was primarily informed by prior advanced research. For instance, Ran et al. (2021) performed a comparative analysis of various methods for ground temperature prediction on the QTP, highlighting the superior performance of SVR. Given the overlap of study area and the higher accuracy of SVR within the context of this study, we were further motivated to adopt this approach. Moreover, SVR provides deterministic results, which makes it especially well-suited for tasks requiring stability and reproducibility, ensuring transparency and consistency in the findings. As observational data continues to accumulate, the deterministic nature of SVR enables a robust evaluation of how new data influences the resulting predictions.

In conclusion, the performance disparities among different predictive methods are clear and can largely be attributed to the inherent characteristics of each algorithm, particularly how each method extracts patterns from the data. Based on a comparative analysis of the results from various methods and informed by previous research, SVR remains the most suitable approach for this study. While other methods similar to RF, such as XGBoost and KNN, could be evaluated, providing an exhaustive comparison within the confines of a single manuscript is challenging. At this stage, further comparisons might risk compromising the coherence and focus of the paper, which is why no additional modifications have been made in the revised manuscript. Nevertheless, we are committed to publishing our responses to your feedback to ensure the transparency of our methodology and potentially offer further insights to the readership. Your comments have highlighted the critical steps in spatial ground temperature prediction, and we plan to conduct a more focused, in-depth analysis of the predictive performance discrepancies between different methods, particularly with respect to the limitations of RF when applied to the existing dataset.

To ensure the manuscript accurately reflects the changes, all modifications have been marked using the track changes feature.

We would like to once again express our sincere gratitude for your valuable feedback!

References:

Nan, Z., Huang, P., and Zhao, L.: Permafrost distribution modeling and depth estimation in the Western Qinghai-Tibet Plateau, Acta. Geogr. Sin., 68, 318-327, 2013 (in Chinese).

Ran, Y., Li, X., Cheng, G., Nan, Z., Che, J., Sheng, Y., Wu, Q., Jin, H., Luo, D., Tang, Z., and Wu, X.: Mapping the permafrost stability on the Tibetan Plateau for 2005–2015, Sci. China Earth Sci., 64, 62–79, doi:10.1007/s11430-020-9685-3, 2021.

Ran, Y., Li, X., Cheng, G., Che, J., Aalto, J., Karjalainen, O., Hjort, J., Luoto, M., Jin, H., Obu, J., Hori, M., Yu, Q., and Chang, X.: New high-resolution estimates of the permafrost thermal state and hydrothermal conditions over the Northern Hemisphere, Earth Syst. Sci. Data, 14, 865–884, doi:10.5194/essd-14-865-2022, 2022.