Dear Reviewer,

Thank you for your effort on review of my submission. Your comments and suggestions are helpful for my current submission and future research. Now, I respond to your comments item-by-item. Your comments in blue and my response in black.

Main comments:

The addition of tests of the same method with fewer parameters and on other measurement sites (SI) is nice and is a significant improvement of the paper. If properly discussed with respect to the main results, it will broaden the conclusions and the applicability of the Transformer_CNN method. Please discuss this (comparison of use with all parameters, restricted set of parameters at QOMS, restricted set of parameters at other site) in the main text, maybe at the end of the Results part.

Thanks.We have added different parameter selections and comparisons of different DL methods with the STEORS site in the revised manuscript.

Table 6:The imputation results for some elements at the QOMS and SETORS sites

Sets	QOMS						SETORS					
	H			LE			H			LE		
	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
Transformer (selected elements)	34.76	4.36	0.74	37.58	4.77	0.69	39.96	5.28	0.70	42.48	4.79	0.67
Transformer_CNN (selected elements)	29.34	3.44	0.83	30.25	3.93	0.78	31.66	4.83	0.80	34.22	4.61	0.79
Transformer_CNN	18.56	3.40	0.95	15.97	2.98	0.89	22.46	3.97	0.91	19.26	3.18	0.86

We all know that many stations on the Tibetan Plateau cannot fully measure the 39 variables listed in the importance ranking of Figure 3. To better validate the model's applicability on the Tibetan Plateau, Table 6 present below the imputation results using Transformer_CNN at the QOMS and SETORS sites (29.77°N, 94.73°E, at an altitude of 4298m), with the year 2012 as the test set. The model employs basic meteorological elements, including single-layer air temperature, pressure, single-layer air humidity, single-layer wind speed, single-layer wind direction, site hourly average precipitation, ground net radiation, single-layer soil temperature, and single-layer soil moisture content. As shown in Table 6, the imputation performance using basic meteorological elements is significantly lower than that using all variables, but it still presents a generally good result. When employing basic meteorological elements for data imputation, the Transformer_CNN model

consistently outperforms the single Transformer model. This superiority is evident at both the QOMS and SETORS sites. Particularly at the SETORS site, the better imputation performance further validates the high applicability of the Transformer_CNN model on the Tibetan Plateau.

The most important issue is the hyper-parameter optimization of DL models. Authors have provided a list of some hyper-parameters in the supplementary materials. But all the parameters have not been listed. For example, the sequence length and number of training epochs for LSTM, GRU, and transformer models has not been listed. More importantly, how did the authors decide that these are the optimal hyper-parameters? Typically, validation data set are used to optimize these parameters. This is really important for a fair comparison between different DL models; otherwise the comparisons are meaningless.

Thanks. In this study, we employed a standardized hyperparameter optimization method. The batch size was set to 32 to control the number of samples used for each parameter update. The learning rate and weight decay parameters were gradually adjusted during training to optimize model performance. The initial learning rate was set at 0.0005 and halved every 6 epochs, with the weight decay parameter set at 0.01. The training was conducted over 100 epochs, with multiple iterations for model training and validation. During each iteration, we recorded the training loss and validation loss, using Mean Squared Error Loss (MSELoss) and Smooth L1 Loss (SmoothL1Loss) to calculate the loss. Gradient clipping (with a maximum norm of 10) was applied to ensure training stability. Specifically, we updated model parameters using the training set and evaluated model performance on the validation set. The primary evaluation metric was the R-squared (R^2) score. The hyperparameter optimization followed these steps:

A) Initial Training: The model was trained with the initial hyperparameter settings and evaluated on the validation set, recording the initial R^2 score on the validation set.

B) Learning Rate Adjustment: After every 6 training epochs, the R^2 score on the validation set was checked. If the score improved, the current model parameter configuration was saved. Otherwise, the learning rate was halved, and training continued.

C) Weight Decay and Batch Size Optimization: Throughout the training process, the weight decay and batch size were kept constant to ensure training stability and convergence.

D) Saving the Best Model: At the end of each training epoch, the R^2 score on the validation set was evaluated. If the current score was better than the previous best score, the current model parameter configuration was saved.

Furthermore, this study did not include sequence length as each sample was treated as an independent feature vector. The model, through convolutional layers and bidirectional recurrent layers (such as LSTM and GRU), is capable of automatically capturing temporal features of the time

series. This approach simplifies the model's complexity while still effectively capturing the characteristics of the time series data. Similar methods have been applied and validated in the literature (Bai et al., 2018; Borovykh et al., 2017; Schuster et al., 1997). By using this approach, we can effectively model time series data without explicitly specifying the sequence length.

Additional comments:

l. 117: higher —> high? the sentence sounds weird, please rephrase

Table 1: please check/homogeneize scientific notation

l. 165: Euclidenian —> Euclidean?

l. 173: I am not sure reliability is the right word here: utility?

Thank you for your valuable feedback. We have addressed your comments and made the necessary revisions in the updated manuscript.

l. 240 and following: the Tibetan plateau is highly variable in its geography, which makes the observation site unique —> discuss in the applicability of the methoh

Section 3: maybe rephrase experiments and methods, with an opening sentence like "we present here the experimental design and the different statistical and learning methods used in this study"

Thank you for your valuable feedback. We have addressed your comments and made the necessary revisions in the updated manuscript.

l. 293: ReLU? it is defined in the SI, please specify

Thank you for your valuable feedback. We have addressed your comments and made the necessary revisions in the updated manuscript.

l. 304: He initialization? reference?

Thank you for your valuable feedback. We have addressed your comments and made the necessary revisions in the updated manuscript.

Figure 4: several terms (GELU, loss functions) are not defined in the main text but in the SI, please make reference to the SI.

Thank you for your valuable feedback. We have addressed your comments and made the necessary revisions in the updated manuscript.

Please check the spelling of citations, throughout the papers (Swinbank and W.C., 1951, Bishop and M 2006)

Thank you for your valuable feedback. We have addressed your comments and made the necessary revisions in the updated manuscript.

However, would it not affect the model validation exercise. A small testing would not give accurate picture of models' capabilities.

We understand your concern regarding the adequacy of model validation given the potentially small test set. To address this issue, we have adopted a rolling testing approach in our study. Each year is sequentially selected as the test set, while the remaining nine years are used for training. This method ensures that the model is evaluated under diverse conditions across different years, providing robust validation.

Notably, due to significant missing turbulence heat flux data in 2012, our paper focuses on data imputation for that year. However, the handling of data for other years is detailed in the supplementary files, ensuring a comprehensive overview of the model's capabilities. This approach leverages a full decade of data, providing thorough validation and a more accurate picture of the model's performance.

How did you find the data with similar environmental conditions?

The lookup table method is based on creating a data retrieval table from valid data, searching for valid data under similar environmental conditions according to major environmental factors, and averaging the found data to impute missing data. Initially, time periods were divided according to six bimonthly intervals or four seasonal periods (April 1 to May 31, June 1 to September 30, October 1 to November 30, and December 1 to March 31). Subsequently, classifications were defined based on air temperature (2°C intervals, ranging from -20°C to 50°C), wind speed (1 m/s intervals, ranging from 0 to 20 m/s), net solar radiation (50 W/m² intervals, ranging from 0 to 1000 W/m²), and relative humidity (10% intervals, ranging from 0% to 100%). For each classification, the mean and standard deviation of turbulent heat flux were compiled. Gaps in the look-up tables were filled using linear interpolation, with a maximum span of 100 W/m² for the net solar radiation curve at given temperatures, 4°C within the air temperature curve at given solar radiation levels, and 20% within the relative humidity curve at given solar radiation and air temperature levels. Finally, the look-up tables were used to find and interpolate the missing turbulent heat flux values based on existing environmental conditions, ensuring the completeness and accuracy of the data.

The validation metrics are very different across the three DL schemes; but the test metrics are not so much. Why?

Thank you for your insightful comment regarding the validation and test metrics across the three deep learning (DL) schemes. The observed differences in validation metrics can be attributed to several factors:

Over fitting to the Training Data: During the training process, each DL scheme might exhibit varying degrees of over fitting to the training data. This over fitting is more apparent in the validation phase, where the model's performance can fluctuate significantly as it tries to generalize from the training data to the validation data.

Differences in Model Complexity: The three DL schemes likely have different architectural complexities. These differences can lead to varying validation performances, as each model architecture responds differently to the same training data.

We sincerely thank you again for your insightful comments, which undoubtedly improved the quality of our work. We have made necessary revisions based on your suggestions regarding the remaining modifications in the PDF, and the improved manuscript now better meets the requirements of the journal.

Best regards,

Sincerely

Quanzhe Hou, Zhiqiu Gao, Zexia Duan, and Minghui Yu
July 15, 2024

References

Bai, S., Kolter, J. Z., and Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271, 2018.

Borovykh, A., Bohte, S., and Oosterlee, C. W.: Conditional time series forecasting with convolutional neural networks, arXiv preprint arXiv:1703.04691, 2017.

Schuster, M., and Paliwal, K. K.: Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing, 45, 2673–2681, https://doi.org/10.1109/78.650093, 1997.