Reply to Reviewer 1

We sincerely thank the reviewer for taking the time to evaluate our manuscript and for providing insightful and constructive comments. We have carefully considered all suggestions and revised the manuscript accordingly. Below, we provide point-by-point responses to the reviewer's comments (responses marked in blue and sections included in manuscript in violet). We fully agree with the reviewer's observations, and all recommended modifications have been incorporated into the revised version of the manuscript.

Specific comments are as follows:

1. The training/testing split is random at the pixel-level, ignoring spatiotemporal autocorrelation. Neighbouring grid cells on the same day or the same cell across days are not independent. Use a more rigorous cross-validation scheme: e.g., leave-one-year-out or spatial block CV (by climate zone or watershed).

Response:

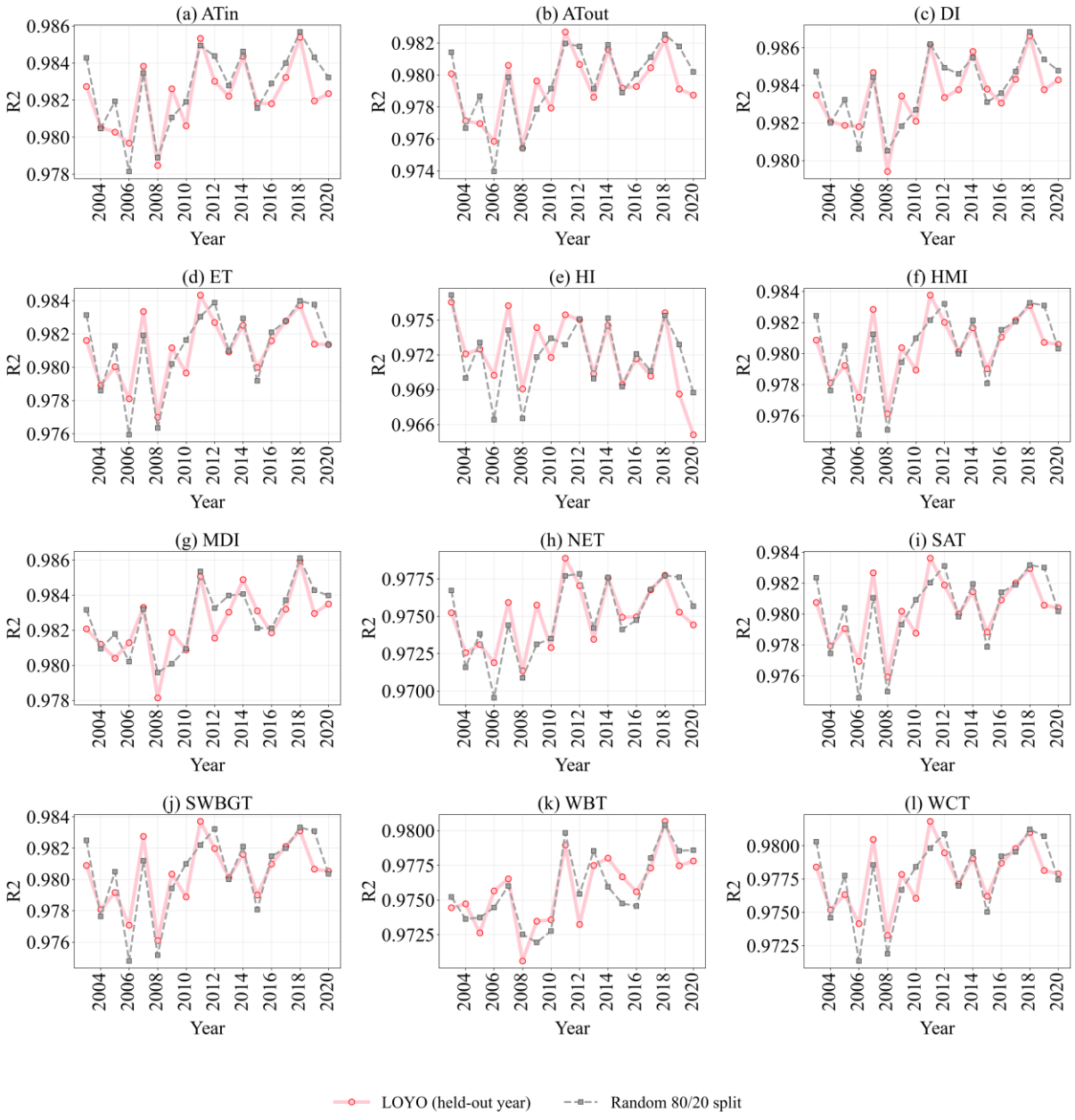
We thank the reviewer for this valuable suggestion. To address the potential influence of spatiotemporal autocorrelation associated with the original random pixel-level train–test split, we implemented Leave-One-Year-Out (LOYO) cross-validation (Bernardo, 1994; Vehtari et al., 2017), in which each year from 2003–2020 was successively withheld for testing while the remaining years were used for model training. This approach provides a more rigorous assessment of temporal generalizability by ensuring complete temporal independence between the training and testing datasets. The LOYO evaluation showed consistently high predictive performance across all 12 HPT indices (R^2 = 0.965–0.987), with only modest year-to-year variations in RMSE and MAE and no systematic decline in any withheld year. Furthermore, the annual accuracies obtained from the LOYO validation closely matched those from the original random 80/20 train–test split (Supplementary Figures 2–4), indicating that the reported model performance is not primarily driven by spatiotemporal autocorrelation between the training and testing samples. Accordingly, we have incorporated the LOYO cross-validation procedure into the Methods section and the corresponding results (lines 150-154) and discussion (lines 456-466) into the Discussion section of the revised manuscript.

Methods:

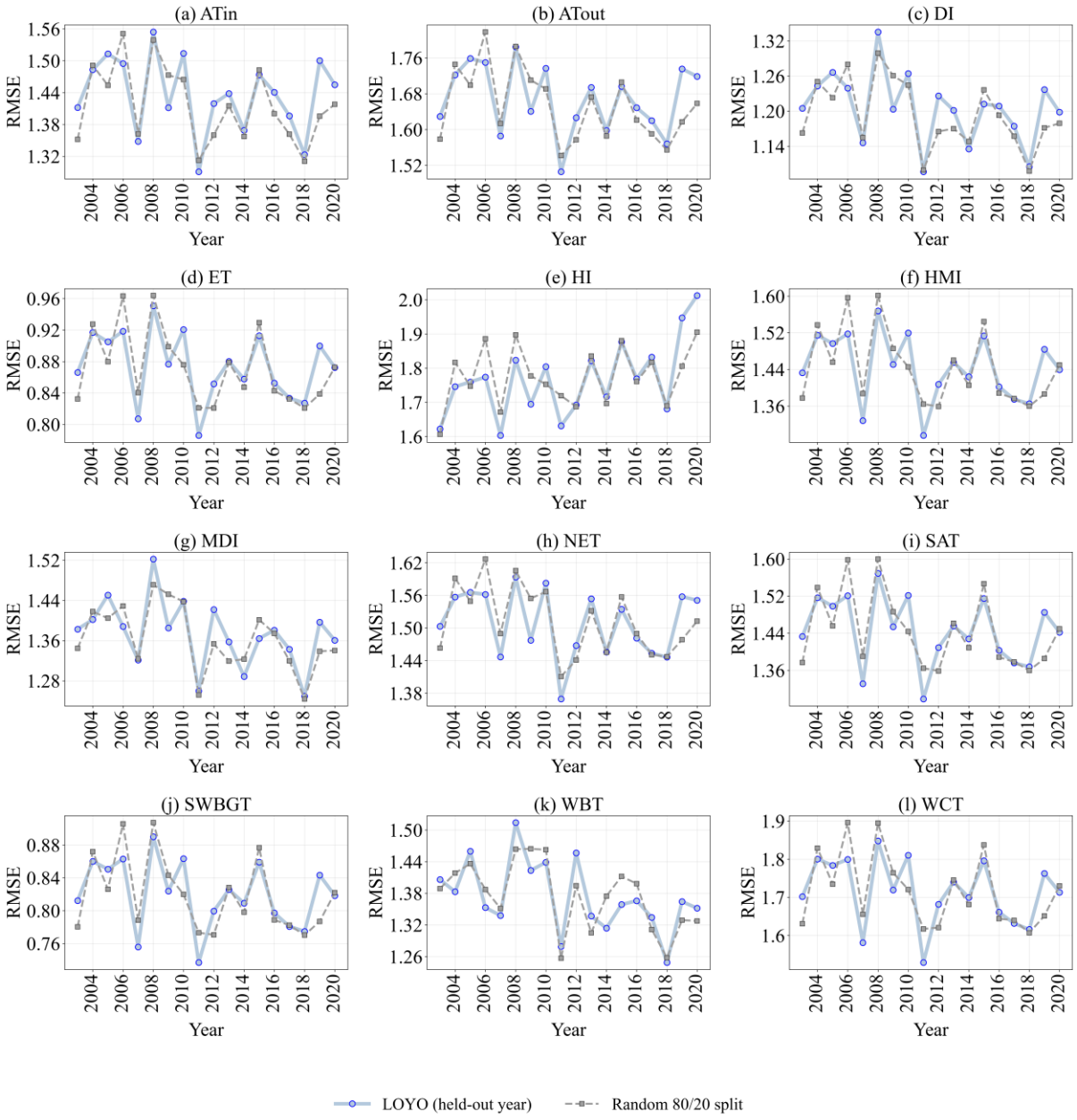
“Also, to examine the influence of spatiotemporal autocorrelation on model performance, a leave-one-year-out validation (Bernardo, 1994; Vehtari et al., 2017) is performed, where each year is withheld once for testing while the remaining years are used for training.”

Discussion:

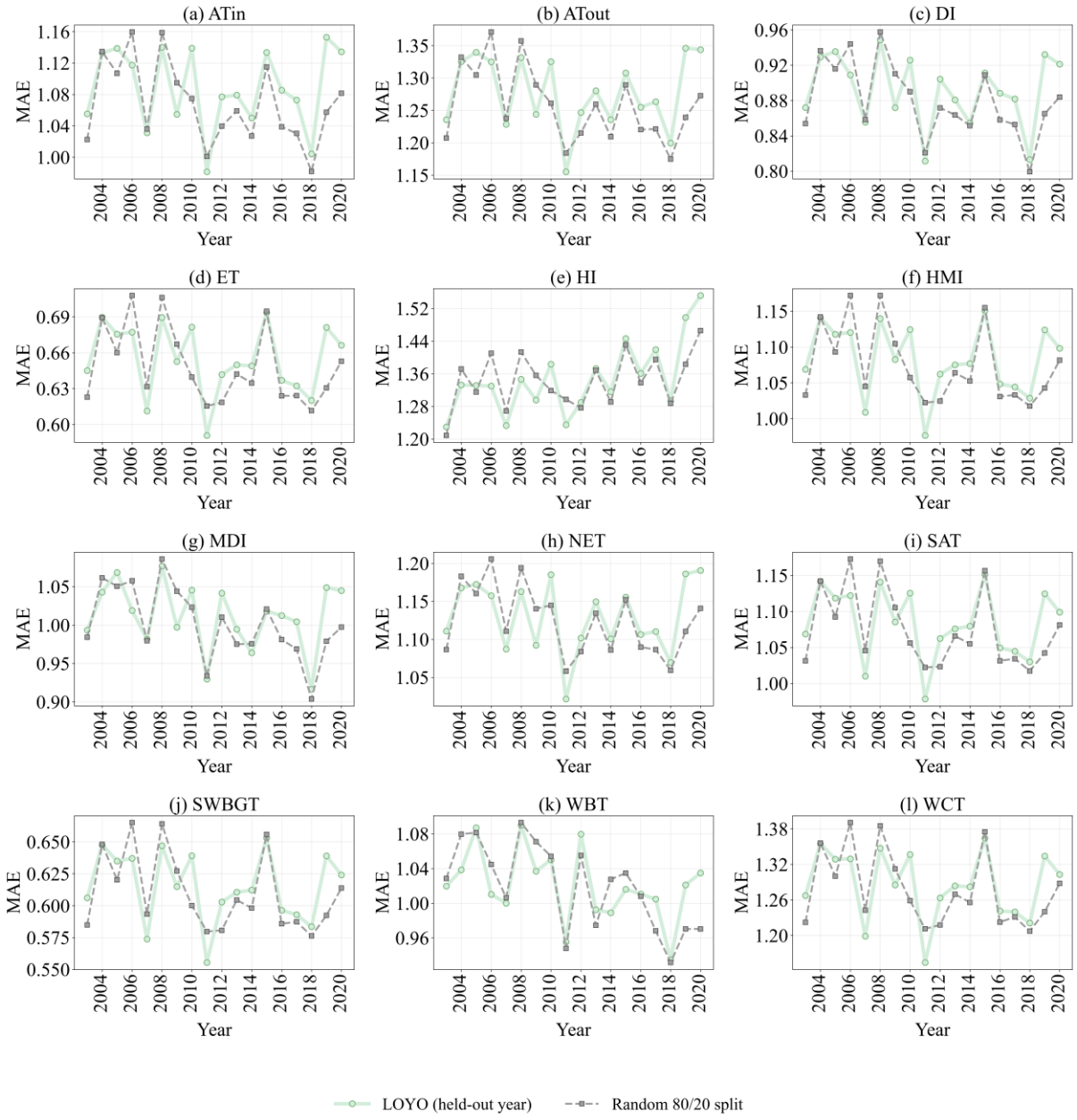
“Additionally, to evaluate whether the reported model accuracy was influenced by spatiotemporal autocorrelation in the random train-test split, we performed LOYO cross-validation (Bernardo, 1994; Vehtari et al., 2017) for all 12 HPTs, successively holding out each year from 2003–2020 for testing. The LOYO performance remained consistently high across all indices (R^2 = 0.965–0.987), with only modest year-to-year variations in RMSE and MAE and no systematic decline in any particular year. Furthermore, the annual accuracies obtained from the LOYO validation closely matched those from the original random 80/20 train–test split (supplementary figures 2–4), indicating that the reported model performance is not primarily driven by spatiotemporal autocorrelation between training and testing samples. These findings demonstrate that the proposed downscaling framework generalizes well across different years and that the reported accuracies are robust to a more rigorous temporal validation strategy.”



"Supplementary Figure 2. Leave-one-year-out (LOYO) cross-validation R^2 for the 12 HPT indices: (a) ATin, (b) ATout, (c) DI, (d) ET, (e) HI, (f) HMI, (g) MDI, (h) NET, (i) SAT, (j) SWBGT, (k) WBT, and (l) WCT. For each held-out year (2003–2020), the model is trained on the remaining 17 years and evaluated on the excluded year. R^2 remains consistently high across all held-out years and indices (approximately 0.965–0.986), with no systematic degradation in any year, indicating stable model performance regardless of the year withheld from training."



"Supplementary Figure 3. RMSE fluctuates within a narrow range across held-out years for each index, with no monotonic trend over the study period, indicating that prediction error is not systematically tied to any specific held-out year."



"Supplementary Figure 4. MAE shows no systematic year-to-year variation across all 12 indices, further supporting the temporal stability of model performance across the 2003–2020 record."

2. All validation is against the same ERA5 data used for training – this only shows how well the model reproduces ERA5, not how accurately the HPT indices represent real human thermal stress. At a minimum, compare model predictions at a few meteorological station locations with station-based HPT calculations using observed T, RH, and wind. If such data are unavailable, explicitly state this as a major limitation and rephrase the dataset as an "ERA5-downscaled" product rather than a "reliable proxy for station observations".

Response:

We thank the reviewer for this valuable suggestion. We agree that validation against independent ground observations is important. Accordingly, we incorporated an independent validation using

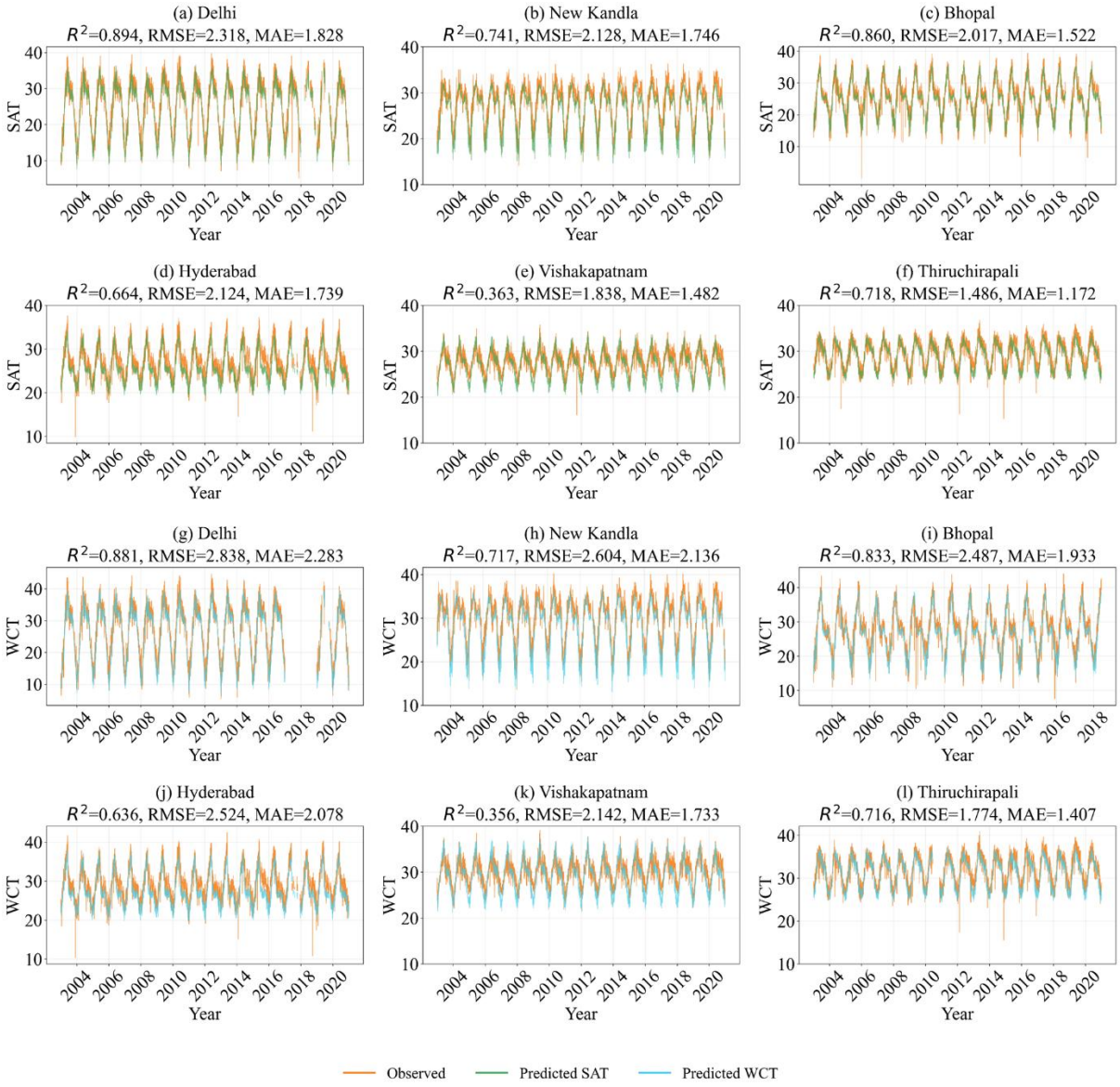
observations from six India Meteorological Department (IMD) stations, where daily maximum air temperature, minimum air temperature, and wind speed data were available. Consequently, independent validation could be performed for SAT and WCT, the two HPT indices that can be directly derived from the available station observations. The validation methodology has been described in the Methods section (line 145-148), and the corresponding results have been added to the Results section (lines 355-366, Figure 10). We also clarify that this comparison provides an independent evaluation of the ERA5-downscaled products against available ground observations, rather than direct validation of all HPT indices.

Method:

"Independent validation of the downscaled products was performed using ground-based observations from six India Meteorological Department (IMD) stations for SAT and WCT, the two HPT indices that could be directly evaluated using the available station measurements."

Results:

"Model predictions for SAT and WCT are validated against independent IMD station observations at six locations. Independent validation was limited to these two HPT indices because only daily maximum air temperature (Tmax), minimum air temperature (Tmin), and wind speed observations were available at the selected IMD stations. Station records are screened for noise by removing values beyond five standard deviations of the station mean, with comparisons restricted to days where both station and predicted values are available. R^2 ranged from 0.363 to 0.894 for SAT (figure 10(a)-10(f)) and 0.356 to 0.881 for WCT across stations (figure 10(g)-10(l)), with the strongest agreement at Delhi and Bhopal and the lowest at Vishakhapatnam. However, at Vishakhapatnam, our predictions closely track the observed seasonal cycle and annual variability, indicating that the model captures the site's overall thermal regime despite greater day-to-day variability. Overall, the comparison demonstrates good agreement between the ERA5-downscaled predictions and the available IMD observations, providing an independent evaluation of the model performance."



"Figure 10. Time series comparison of predicted and observed (a–f) surface air temperature (SAT) and (g–l) wind chill temperature (WCT) at six India Meteorological Department (IMD) stations (Delhi, New Kandla, Bhopal, Hyderabad, Visakhapatnam, and Thiruchirapalli) during 2003–2020. Observed SAT was derived as the daily mean of the maximum and minimum air temperatures, whereas observed WCT was calculated from the daily mean air temperature and wind speed using the same equation applied to the HiTIC-India dataset. Orange lines denote the in-situ observations, while green and cyan lines denote the predicted SAT and WCT, respectively."

3. The inclusion of LST, PWV, population, slope, aspect, and elevation is plausible but not analysed. LST is highly correlated with air temperature and might dominate the model. Perform feature importance analysis (e.g., SHAP) and report variance inflation factors.

Response:

We thank the reviewer for this valuable suggestion. To evaluate whether LST, because of its strong correlation with air temperature, disproportionately influenced the model predictions, we performed a SHAP (SHapley Additive exPlanations) feature importance analysis (Lundberg and Lee, 2017) for all 12 HPT indices. The analysis showed that elevation and day of year were consistently the two most influential predictors, together accounting for approximately 55–65% of the total feature importance, whereas LST contributed only 5–18%. PWV, latitude, and population density also provided meaningful contributions, while slope, aspect, and climate zone had comparatively lower importance. In addition, we assessed multicollinearity among all predictors using the Variance Inflation Factor (VIF) (Kutner et al., 1996; Thompson et al., 2017). All VIF values were below 3.2, well under the commonly accepted threshold of 5, indicating that the selected covariates provide complementary and largely independent information. These results demonstrate that the downscaling framework is not dominated by LST alone but is driven by a combination of climatic, topographic, and demographic factors. Accordingly, we have incorporated the SHAP feature importance and VIF analyses into the revised manuscript in Methods (lines 156-159), and discussed their implications in the Discussion section (lines 430-448). The corresponding results are presented in Figures 14 and 15.

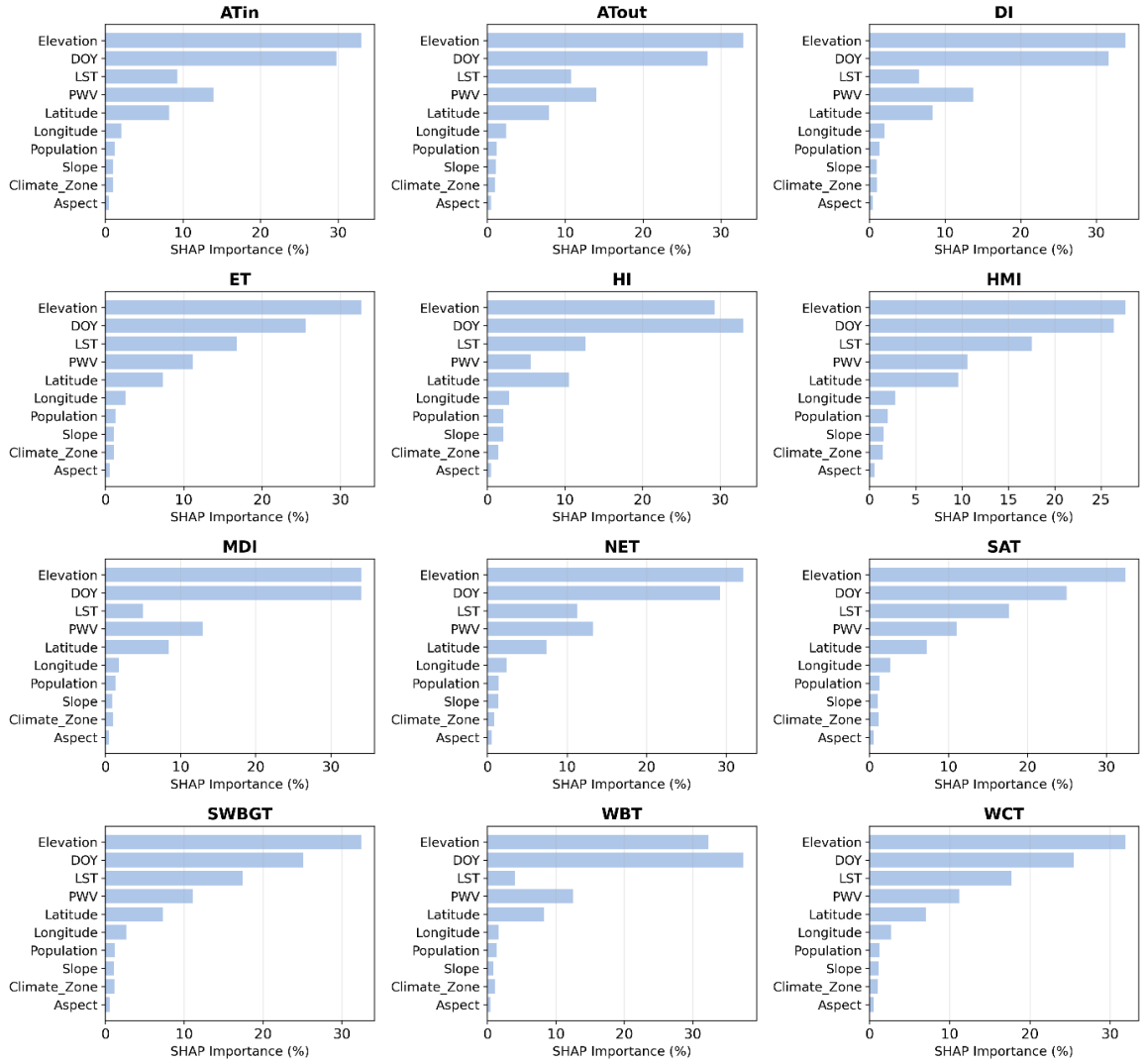
Method:

"Additionally, Shapley Additive explanations (SHAP) analysis (Lundberg and Lee, 2017)and the Variance Inflation Factor (VIF) are employed to quantify the contribution of each covariate to the HPT predictions and to assess multicollinearity among the predictor variables."

Discussion:

"Understanding the relative influence of predictor variables is essential for interpreting the physical basis of machine learning models used for thermal downscaling. SHAP analysis (Lundberg and Lee, 2017) showed that elevation and doy were consistently the two most influential predictors across all 12 HPTs, together accounting for approximately 55–65% of the total feature importance (figure 14). This indicates that topographic and doy (seasonal controls) are the primary drivers of thermal variability

across India. While lst contributed only about 5–18%, pwv, latitude, and population density provided smaller but still meaningful contributions (2–14%). Aspect, slope, and climatic zones showed consistently low importance across the indices (figure 14). Overall, the SHAP results indicate that the downscaling framework captures the complex interactions among topographic, climatic, and surface-related variables that govern human thermal conditions."



"Figure 14. SHAP feature importance (%) for the 10 covariates used in the LightGBM models for each of the 12 HPT indices."

"The Variance Inflation Factor (VIF) analysis further demonstrated that the covariates set did not exhibit multicollinearity, with all VIF values remaining below 3.2 (figure 15), and well under the commonly used threshold of 5 (Kutner, 1996; Thompson et al., 2017). The highest value was associated with Elevation, which is expected because of its natural correlation with slope and temperature-related

variables (LST). These results indicate that LST, pwv, population density, and the topographic covariates provide complementary and largely independent information, supporting the physical consistency and robustness of our downscaling framework."

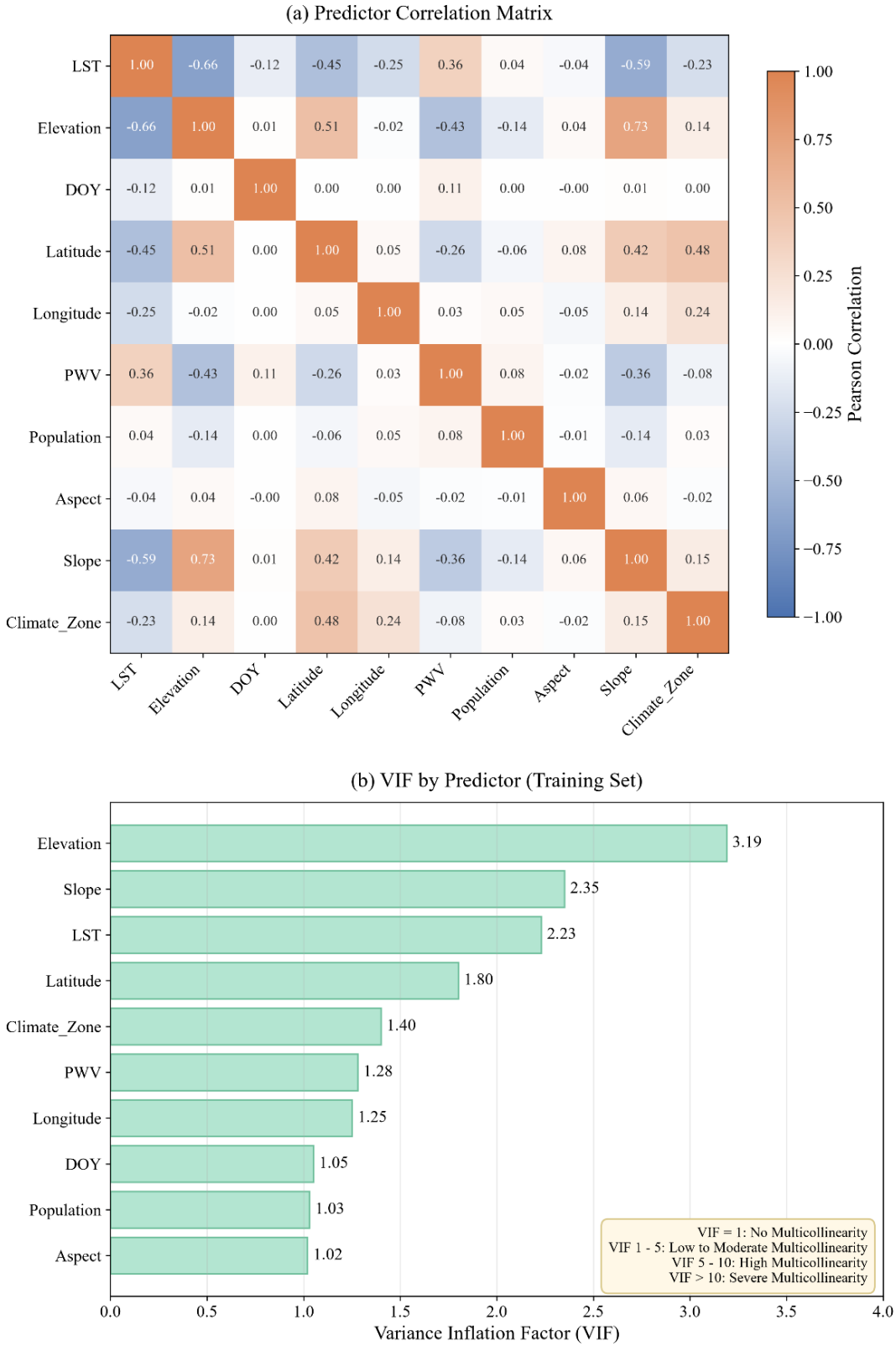


Figure 15. Multicollinearity assessment of the 10 predictor variables used in the LightGBM models. (a) Pearson correlation matrix among predictors. (b) Variance inflation factor (VIF) for each predictor, computed on the training set. All predictors fall well below the conventional multicollinearity threshold of 5. The absence of severe multicollinearity (VIF > 10) indicates that LST, PWV, population, and the topographic covariates each contribute largely independent, complementary information to the model.

4. The manuscript only mentions "grid search with 5-fold CV". Provide a table of the hyperparameter search space and the final chosen values (e.g., number of leaves, learning rate, boosting iterations, min data in leaf, feature fraction). Without this, the results are not reproducible.

Response:

We thank the reviewer for this valuable suggestion. To improve the reproducibility of our modelling framework, we have expanded the description of the hyperparameter optimization procedure in the Methods section. Specifically, we now describe that hyperparameter optimization was performed using a grid search strategy (Bergstra et al., 2012) with 5-fold cross-validation, systematically exploring predefined parameter ranges for tree structure (number of leaves and maximum depth), leaf-level regularization, row and feature sampling fractions, L1/L2 regularization, learning rate, and categorical feature handling. The optimal parameter combination for each of the 12 HPT models was selected based on the best mean cross-validation performance (A Ilemobayo et al., 2024; Ngoc et al., 2022). In addition, we have included the complete hyperparameter search space and the final selected values for all 12 HPT indices in Supplementary Table S1, thereby ensuring full reproducibility of the reported result and updated in the methods section (lines 127-134).

Method:

"Hyperparameter optimization is performed using a grid search strategy (Bergstra et al., 2012)based on 5-fold cross-validation, systematically exploring predefined parameter ranges governing tree structure (number of leaves, maximum depth), leaf-level regularization, row and feature sampling fractions, L1/L2 regularization strength, learning rate, and categorical feature handling. For each candidate configuration, the model was iteratively trained and evaluated across all five folds, with the configuration yielding the best mean validation performance selected for each of the 12 HPTs (A

Ilemobayo et al., 2024; Ngoc et al., 2022). The hyperparameter search space and the chosen values are provided in supplementary table 1.”

“Supplementary Table 1. Hyperparameter search space and final tuned values for the LightGBM models of the 12 HPT indices, obtained via grid search across six categories: tree complexity (num_leaves, max_depth), leaf regularization (min_data_in_leaf, min_sum_hessian_in_leaf), row/feature sampling (feature_fraction, bagging_fraction), L1/L2 regularization (lambda_l1, lambda_l2), learning rate and split gain (learning_rate, min_gain_to_split), and categorical regularization (cat_smooth, cat_l2). The search space column lists the candidate values evaluated for each parameter, and the index-specific columns report the final selected value for each of the 12 HPT indices.”

Sl no.	Name	Parameter Tuned	Search_Space	ATin	ATout	DI	ET	HI	HMI	MDI	NET	SAT	SWBGT	WBT	WCT
1	Tree complexity	num_leaves	{31, 63, 127, 255}	255	255	255	255	255	255	255	255	255	255	255	255
		max_depth	{-1, 8, 12, 16}	-1	16	16	-1	16	16	16	-1	-1	-1	16	-1
2	Leaf regularization	min_data_in_leaf	{20, 50, 100, 200}	50	50	50	50	50	50	20	50	50	50	50	50
		min_sum_hessian_in_leaf	{0.001, 0.1, 1.0}	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3	Row/feature sampling	feature_fraction	{0.70, 0.85, 1.00}	0.85	0.85	0.85	0.85	0.7	0.7	0.85	0.85	0.85	0.85	0.85	0.85
		bagging_fraction	{0.70, 0.85, 1.00}	0.85	0.7	0.85	0.85	0.85	0.85	0.85	0.7	0.85	0.85	0.85	0.85
4	L1/L2 regularization	lambda_l1	{0.0, 0.1, 1.0, 5.0}	5	5	1	0	5	5	5	1	5	1	0	0
		lambda_l2	{0.0, 0.1, 1.0, 5.0}	5	1	5	5	0	0	1	0	0.1	1	5	5
5	Learning rate & split gain	learning_rate	{0.02, 0.05, 0.08, 0.10}	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
		min_gain_to_split	{0.0, 0.05, 0.20}	0.05	0.05	0.2	0	0.05	0.2	0.2	0.2	0	0	0	0.2
6	Categorical regularization	cat_smooth	{1.0, 10.0, 30.0}	30	30	30	30	1	10	1	10	10	1	10	10
		cat_l2	{1.0, 10.0, 30.0}	1	1	1	1	30	10	1	1	1	1	1	1