Supporting Information for "Forecasting Tropical Annual Maximum Wet-Bulb Temperatures Months in Advance from the Current State of El Niño"

Yi Zhang[1,2], William R. Boos[1,3], Isaac Held[4], Christopher J. Paciorek[5],

Stephan Fueglistaler[4,6]

[1]Department of Earth and Planetary Science, University of California, Berkeley, CA 94720

[2]Miller Institute for Basic Research in Science, University of California, Berkeley, CA 94720

[3]Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, CA 94720

[4]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ 08540

[5]Department of Statistics, University of California, Berkeley, CA 94720

[6]Department of Geosciences, Princeton University, Princeton, NJ 08540

Contents of this file

Text S1. Fitting generalized extreme value distributions

The generalized extreme value (GEV) distribution is represented by the probability density function:

$$f(x;\mu,\sigma,\xi) = \begin{cases} \exp\left(-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right)\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-(\frac{1}{\xi}+1)}\frac{1}{\sigma}, & \text{for } \xi \neq 0, \\ \exp\left(-\exp\left(\frac{-(x-\mu)}{\sigma}\right)\right)\exp\left(\frac{-(x-\mu)}{\sigma}\right)\frac{1}{\sigma}, & \text{for } \xi = 0. \end{cases} \tag{1}$$

In our model, the location parameter, μ, is defined as a linear combination of the year (t) and the Oceanic Niño Index from the preceding December (ONI_{t-1}):

$$\mu = \beta_0 + \beta_1 t + \beta_2 \mathrm{ONI}_{t-1} \tag{2}$$

Thus, the log-likelihood function, given the data, can be expressed as:

$$\mathrm{L}(\beta_0, \beta_1, \beta_2, \sigma, \xi | \mathrm{TW}_{\mathrm{max},1}, ...\mathrm{TW}_{\mathrm{max,n}}, t_1, .., t_n, \mathrm{ONI}_0, ..., \mathrm{ONI}_{t-1})$$

$$= \begin{cases} -n\ln\sigma - (1+\frac{1}{\xi})\sum_{t=1}^{n}\ln(1+\xi\frac{\mathrm{TW}_{\mathrm{max}}-\beta_0-\beta_1 t-\beta_2\mathrm{ONI}_{t-1}}{\sigma}) - \sum_{t=1}^{n}(1+\xi\frac{\mathrm{TW}_{\mathrm{max}}-\beta_0-\beta_1 t-\beta_2\mathrm{ONI}_{t-1}}{\sigma})^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ -n\ln\sigma - \sum_{t=1}^{n}\frac{\mathrm{TW}_{\mathrm{max}}-\beta_0-\beta_1 t-\beta_2\mathrm{ONI}_{t-1}}{\sigma} - \sum_{t=1}^{n}\exp(-\frac{\mathrm{TW}_{\mathrm{max}}-\beta_0-\beta_1 t-\beta_2\mathrm{ONI}_{t-1}}{\sigma}) & \text{if } \xi = 0 \end{cases}$$

To obtain optimal parameter values, the negative log-likelihood ($-\mathrm{L}$) is minimized using the "minimize" function from Python's "scipy" package. The estimated $\hat{\beta}$ values derived this way, as well as the Mean Squared Error (MSE), are quite similar to those from a standard multiple linear regression (not shown).

Text S2. Cross-validation of the regression model

Before generating predictions with the multiple linear regression model, we evaluate its predictive performance through leave-one-out cross-validation and walk-forward validation.

The leave-one-out cross-validation results in a 9.4% increase in the Root Mean Squared Error (RMSE) for the tropical mean. Similarly, the increase in RMSE ranges from 8.9% to 9.7% for each of the four focus regions, as shown in Table 1.

For the walk-forward validation, we use the first 20 data points as the training set and progressively add one data point to the training set. The validation set comprises the one data point immediately following the training set. The results are presented in Figure S4 and Table 1. The RMSE for the last 19 data points shows a 7.1% increase for the tropical mean and ranges from 17.6% to 20.1% for the four regions.

These moderate increases in RMSE from both the leave-one-out cross-validation and walk-forward validation suggest that the model is not seriously overfitted.

Text S3. Confidence intervals and prediction intervals

The 95% confidence interval of the predicted mean $\hat{\mathrm{TW}}_{\mathrm{max}}$ is calculated as:

$$\hat{\mathrm{TW}}_{\mathrm{max}} \pm t_{0.025,36}\ \mathrm{se}(\hat{\mathrm{TW}}_{\mathrm{max}}) \tag{3}$$

where the t value of a two-sided significance level of 0.05 with 36 degrees of freedom is 2.028. The 95% prediction interval is given by:

$$\hat{\mathrm{TW}}_{\mathrm{max}} \pm t_{0.025,36}\ \sqrt{\mathrm{MSE} + \mathrm{se}(\hat{\mathrm{TW}}_{\mathrm{max}})^2}, \tag{4}$$

where the standard error of fit at $\mathbf{x_p} = (1, 2024, \mathrm{ONI}_{\mathrm{Dec,2023}})^{\mathrm{T}}$ is determined by

$$\mathrm{se}(\hat{\mathrm{TW}}_{\mathrm{max}}) = \sqrt{\mathrm{MSE}(\mathbf{x_p}^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}\mathbf{x_p})} \tag{5}$$

with MSE being the mean squared error and X being the 39×3 regressor matrix:

$$X = \begin{bmatrix} 1 & 1980 & \mathrm{ONI}_{\mathrm{Dec},1979} \\ 1 & 1981 & \mathrm{ONI}_{\mathrm{Dec},1980} \\ 1 & 1982 & \mathrm{ONI}_{\mathrm{Dec},1981} \\ 1 & 1985 & \mathrm{ONI}_{\mathrm{Dec},1984} \\ \vdots & \vdots & \vdots \\ 1 & 1991 & \mathrm{ONI}_{\mathrm{Dec},1990} \\ 1 & 1994 & \mathrm{ONI}_{\mathrm{Dec},1993} \\ \vdots & \vdots & \vdots \\ 1 & 2022 & \mathrm{ONI}_{\mathrm{Dec},2021} \end{bmatrix}. \tag{6}$$

Text S4. Confidence intervals for the exceedance probability

We calculate the confidence intervals of the probability by bootstrapping the original 39 data points and associated covariate values (a nonparametric bootstrap). For each of the 9999 bootstrap replicates, we estimate the probability of exceeding the previous record using the multiple linear regression model. Specifically, we assume that the predicted $\mathrm{TW}_{\mathrm{max}}$ at each $\mathbf{x_p} = (1, 2024, \mathrm{ONI}_{\mathrm{Dec},2023})^{\mathrm{T}}$ follows a Gaussian distribution centering at the predicted mean given by Equation (6) with a standard deviation equalling the root mean squared error (RMSE). For each $\mathrm{ONI}_{\mathrm{Dec},2023}$ value, we then compute the area under this Gaussian distribution when the predicted $\mathrm{TW}_{\mathrm{max}}$ exceeds the highest record, resulting in the probability of a new $\mathrm{TW}_{\mathrm{max}}$ record being set for this replicate. Given the potential skewness of the probability distribution when the central estimate approaches 0 or 1, we employ the bias-corrected and accelerated (BCa) method for bootstrap confidence intervals. The BCa bootstrap analysis is conducted using the R “boot” package.

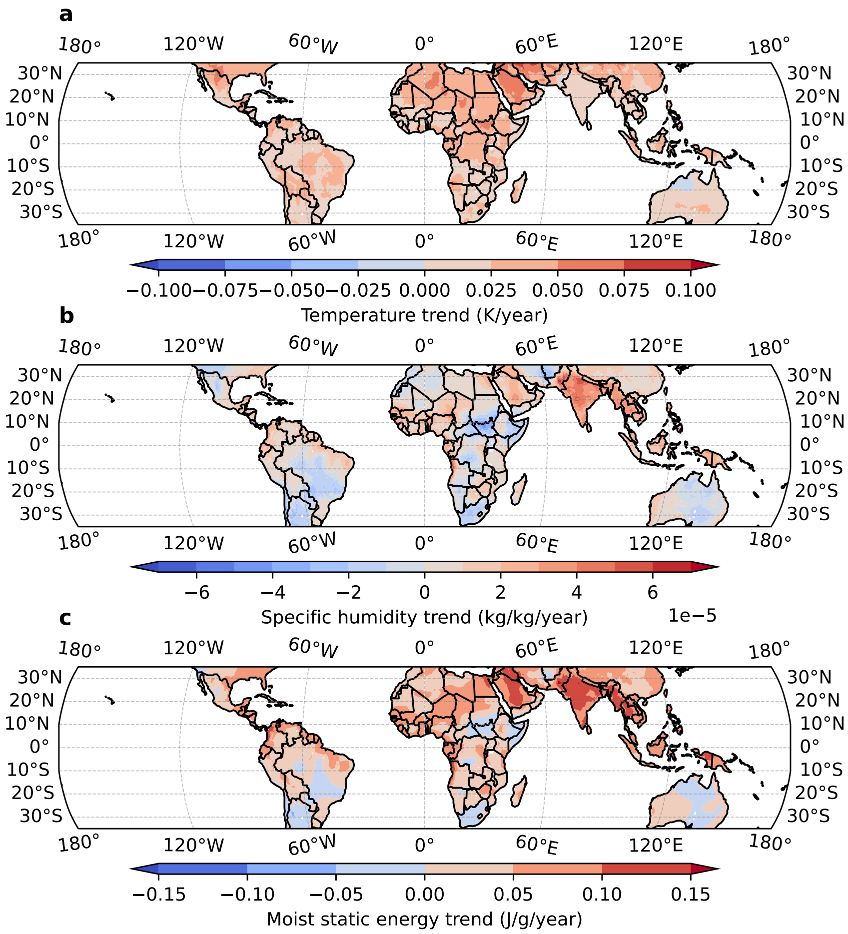

Figure S1. Linear trends of annual-mean 2-m air temperature (**a**), specific humidity (**b**), and moist static energy (**c**) from 1979 to 2022 according to ERA5.

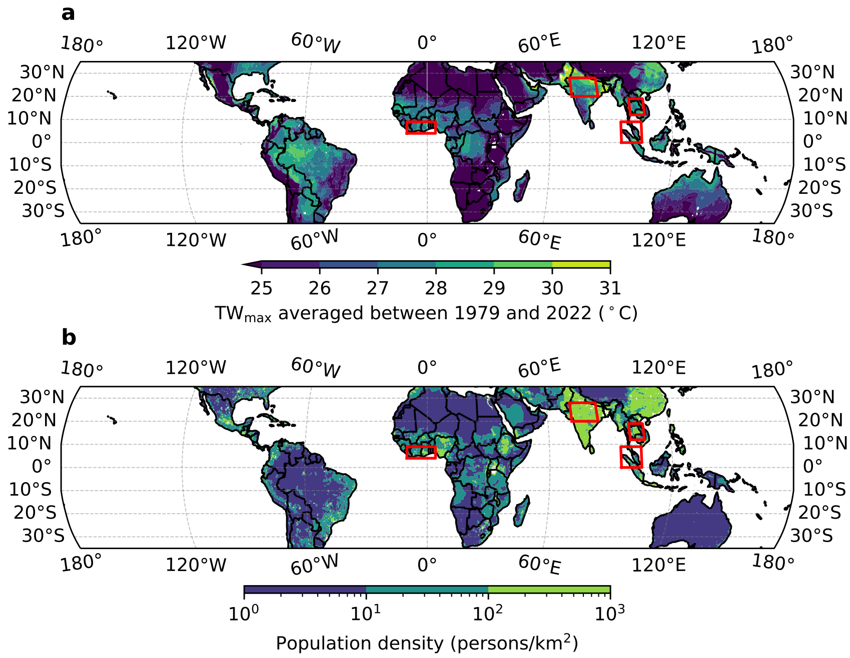

Figure S2. **a**, Climatology of the annual maximum TW (TW_{max}) between 1979 and 2022. The color scale only distinguishes values greater than 25°C. **b**, Population density map acquired from `https://neo.gsfc.nasa.gov/view.php?datasetId=SEDAC_POP`.

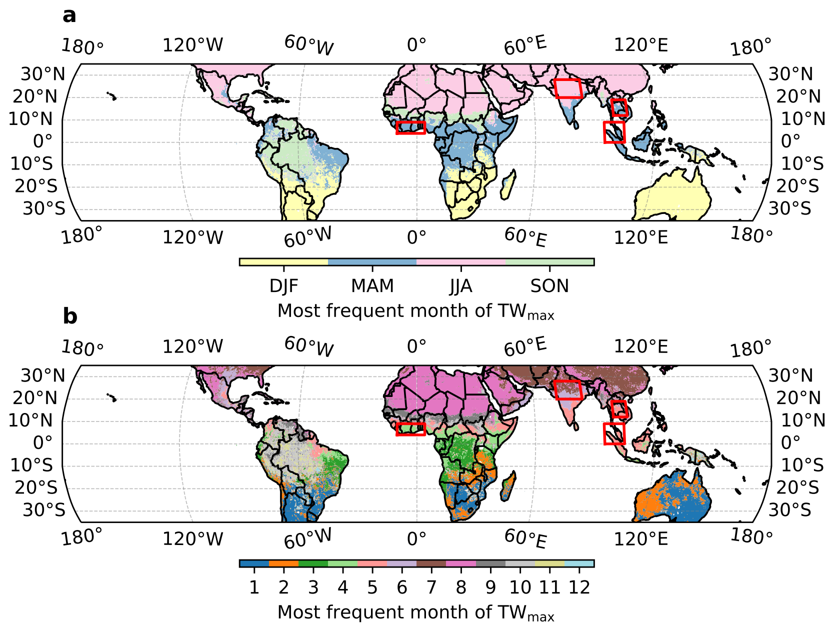

Figure S3. The most common month for the occurrence of annual maximum TW for each location. Groupings of three months (DJF, MAM, JJA, SON) are shown in **b**, along with individual months.

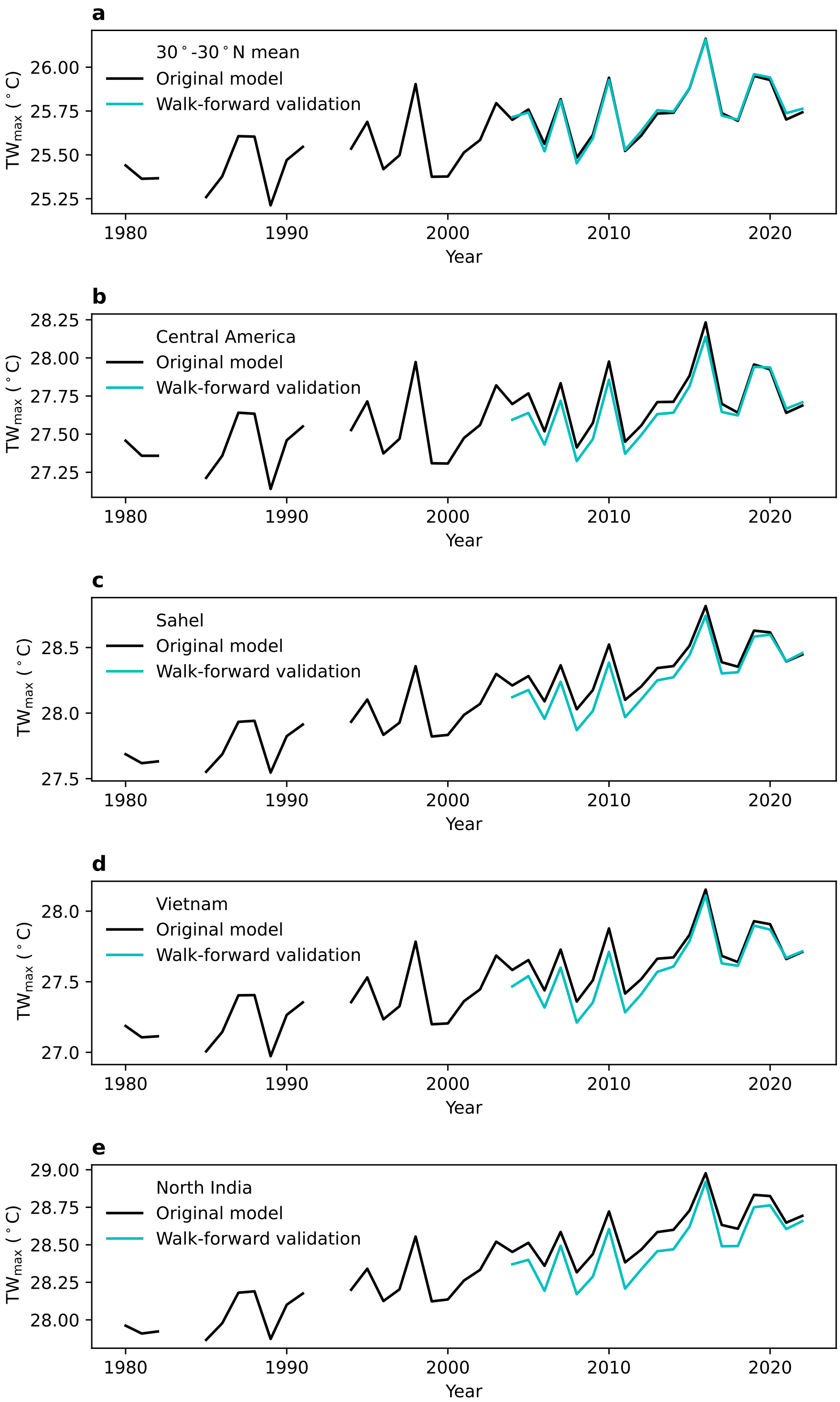

Figure S4. Time series of predicted TW_{max} from the original model with all the 39 data points and that from walk-forward validation for the 30°S-30°N land average (**a**) and for each of the four defined regions of interest (**b-e**). The initial training set contains the first 20 data points. The missing points are the post-volcanic eruption years.