Supplementary information for "Daytime-only-mean data can enhance understanding of land-atmosphere coupling"

Zun Yin[1], Kirsten Findell[2], Paul Dirmeyer[3], Elena Shevliakova[2], Sergey Malyshev[2], Khaled Ghannam[1], Nina Raoult[4], and Zhihong Tan[1]

[1]Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, 08540, New Jersey, USA
[2]Geophysical Fluid Dynamics Laboratory, NOAA/OAR, Princeton, 08540, New Jersey, USA
[3]Center for Ocean-Land-Atmosphere Studies, George Mason University, Fairfax, 22030, Virginia, USA
[4]Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CNRS-CEA-UVSQ, Gif-sur-Yvette, 91191, Essonne, France

Correspondence: Zun Yin (zyin@princeton.edu)

This PDF file includes:

- Supplementary text 1 "The key driver of L-A coupling signal attenuation due to monthly smoothing"
- Supplementary text 2 "Atmospheric advection-dominated climate regime in Sahara"
- Figures S1 to S7

Supporting Information Text

The key driver of L-A coupling signal attenuation due to monthly smoothing

First, we introduce the algorithms of both trend and seasonal cycle removal applied to the original time series. Then, we check that the detrended-seasonal removed monthly time series is equal to the monthly mean of the detrended-seasonal removed daily time series. Finally, we separate the two-legged metrics (TLM) into the standard deviation term (σ) and the correlation coefficient term (ρ), and investigate the key factor leading to the difference between monthly- and entire-day-mean-based TLM.

Detrending and removal of the seasonal cycle. Let's consider a daily time series x_i. To calculate the two-legged metrics, both trend and seasonality must be removed from the original values. To remove the long-term trend, we generate a linear regression model between time and the variable of interest (e.g., x_i), and then perform detrending by removing model-predicted values from original values like

$$\dot{x}_i = x_i - g(i) \tag{S1}$$

where i is day index and $\dot{x}_i$ is detrended time series. $g(i)$ is the linear regression function retrieved from the x_i against time.

To remove the seasonal cycle, we estimate the seasonality by calculating the multi-year mean of the target value at a specific date, and then perform the removal as

$$\tilde{x}_{d,y} = \dot{x}_{d,y} - \frac{1}{Y}\sum_{i=1}^{Y}\dot{x}_{d,i} \tag{S2}$$

where $\tilde{x}_{d,y}$ is the time series after removing the seasonality. The subscript (d, y) represents time in the form of date and year, and Y is the number of years in the averaging.

Daily and monthly time series. Here we demonstrate that detrended-seasonal removed monthly time series is equal to the monthly mean of detrended-seasonal removed daily time series. Let's assume a detrended daily time series data o_t ($t \in [1, D \times M \times Y]$). Here D, M, and Y are the numbers of day in a month, the number of months, and the number of years, respectively. The time step t can be written in the form of {day,month,year} as $t = \{d, m, y\}$ ($d \in [1, D]$, $m \in [1, M]$, $y \in [1, Y]$).Then we can get the seasonal removed daily time series O_t as

$$O_{d,m,y} = o_{d,m,y} - \frac{1}{Y}\sum_{k=1}^{Y} o_{d,m,k} \tag{S3}$$

The detrended monthly time series p_t (t can be written as $\{m, y\}$) is

$$p_{m,y} = \frac{1}{D}\sum_{i=1}^{D} o_{i,m,y} \tag{S4}$$

The seasonal removed monthly time series P_t is

$$\begin{aligned}
P_{m,y} &= p_{m,y} - \frac{1}{Y}\sum_{k=1}^{Y} p_{j,k} \\
&= \frac{1}{D}\sum_{i=1}^{D} o_{i,m,y} - \frac{1}{Y}\sum_{k=1}^{Y} p_{m,k} \\
&= \frac{1}{D}\left(\sum_{i=1}^{D} o_{i,m,y} - \frac{1}{Y}\sum_{i=1}^{D}\sum_{k=1}^{Y} o_{i,m,k}\right) \\
&= \frac{1}{D}\sum_{i=1}^{D}\left(o_{i,m,y} - \frac{1}{Y}\sum_{k=1}^{Y} o_{i,m,k}\right) \\
&= \frac{1}{D}\sum_{i=1}^{D} O_{i,m,y}
\end{aligned} \tag{S5}$$

Differences between M- and E-based TLMs. First, let's have a look at the σ term of the TLMs. To keep the symbols simple, we denote a_i and b_i (i is day index) as detrended and seasonal removed daily time series. A_j and B_j (j is the month

index) are corresponding monthly time series. As the long-term average of b_i (i.e., $\bar{b}$) is zero, the σ_b can be expressed as

$$\begin{aligned}\sigma_b &= \left(\frac{1}{DMY}\sum_{i=1}^{DMY} b_i^2 - \bar{b}^2\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{MY}\sum_{i=1}^{MY}\left(\frac{b_i^2 + b_{i+1}^2 + b_{i+2}^2 + ... + b_{i+D}^2}{D}\right)_j\right)^{\frac{1}{2}}\end{aligned} \tag{S6}$$

D, M, and Y are the number of days, months, and years, respectively. The σ_B can be written as

$$\begin{aligned}\sigma_{B_j} &= \left(\frac{1}{MY}\sum_{j=1}^{MY} B_j^2 - \bar{B}^2\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{MY}\sum_{j=1}^{MY}\left(\frac{\sum_{i\in j}^{D} b_i}{D}\right)^2\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{MY}\sum_{j=1}^{MY}\left[\frac{\left(b_i + b_{i+1} + b_{i+2} + ... + b_{i+D}\right)^2}{D^2}\right]_j\right)^{\frac{1}{2}}\end{aligned} \tag{S7}$$

The difference between σ_b and σ_B is illustrated in Fig. S1. σ_b contains all squared b_i (dark boxes in Fig. S1), but σ_B contains averaged products of all combinations of b_i within a month.

It is not difficult to proof that $D^2\sum_{i=1}^{N} b_i^2 \geq \left(b_i + b_{i+1} + ... + b_N\right)^2$. The equal relation stands when $b_i = b_{i+1} = ... = b_N$, indicating all daily variables are the same within a month. Considering all months, the σ_B is larger if b_i follows the Matthew principle better, that is large values assemble together in specific months and small values assemble together in other months. As b_i is a time series of variables in a natural process. b_i is somehow correlated with itself at a certain time scale, that is the memory of b_i. It implies that if b_i is large, its neighbours (e.g., b_{i-1} and b_{i+1}) are large as well. Thus, the memory (characterized by auto-correlation) may determine the information loss from σ_b to σ_B, if the σ_b is considered as the accurate information we want.

The ρ term based on daily time series can be written as:

$$\begin{aligned}\rho(a,b) &= \frac{\sum_{i=1}^{DMY}\left(a_i - \bar{a}\right)\left(b_i - \bar{b}\right)}{\sigma_a\sigma_b} \\ &= \frac{\sum_{i=1}^{DMY} a_i b_i}{\sigma_a\sigma_b}.\end{aligned} \tag{S8}$$

$\bar{a}$ and $\bar{b}$ are mean of a_i and b_i, respectively. Similarly, we can get $\rho(A,B)$ as

$$\begin{aligned}\rho(A,B) &= \frac{\sum_{j=1}^{MY}\left(A_j - \bar{A}\right)\left(B_j - \bar{B}\right)}{\sigma_A \cdot \sigma_B} \\ &= \frac{1}{\sigma_A\sigma_B}\sum_{j=1}^{MY}\left(\frac{\left(\sum_{i\in j} a_i\right)\left(\sum_{i\in j} b_i\right)}{D^2}\right).\end{aligned} \tag{S9}$$

The ρ term contains σ terms, which has been discussed in the previous section. If we focus on the numerator, we can find that the difference of numerator between E and M has a similar structure as the ρ difference between E and M. Thus, we deduct that the cross-covariance between a_i and b_i is the key contributor to the difference of the ρ's numerator between E and M.

Atmosphric advection-dominated climate regime in Sahara

Unlike most other places, the atmospheric leg ($\mathcal{A}$) across the Sahara region is negative (Fig. S4), suggesting a negative correlation between the sensible heat flux (H) and the pressure at the LCL (P_{lcl}). This atypical signal is present in all seasons and may be caused by a special mechanism driven by atmospheric advection. Northerly winds from the Mediterranean Sea cool and moisten the near-surface air of the Sahara region, while southerly winds warm and dry the surface (Fig. S6a). According to ERA5, the correlation between E-based daily northward wind speed ($v_{10\text{m}}$) and the 2-m air temperature ($T_{2\text{m}}$) for ten-year JJA data at a sample grid cell in the Sahara is 0.63 (Fig. S6b), which is much larger than that of the eastward wind case (0.12, not shown). On the other hand, the northerly winds show a high correlation with the 2-m absolute humidity (AH), as well (-0.67, Fig. S6b). This suggests that atmospheric advection may determine the inter-daily fluctuations of near-surface temperature and humidity rather than the sensible heat flux from the surface. One piece of evidence is that $T_{2\text{m}}$ fluctuates synchronously with H in the Sahara, with Fig. S6c showing that the auto-correlation is strongest with no time lag between variables. If the $T_{2\text{m}}$ is driven by the surface through H then the peak correlation should occur with a few hours time lag between H and $T_{2\text{m}}$, as shown for an example European grid cell in Fig. S6d.

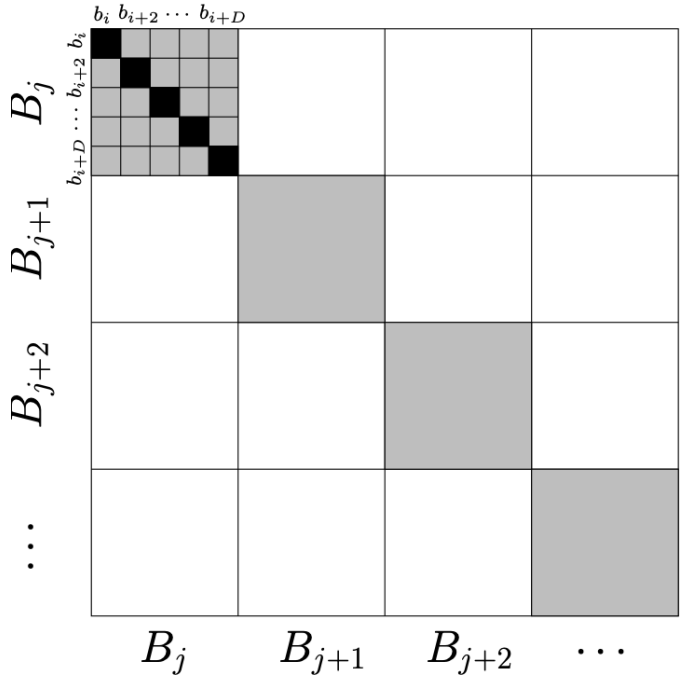

Figure S1. Illustration of the difference between σ_b and σ_B. Small boxes indicate daily time series of b_i. And large boxes indicate monthly time series B_j. For month j (i.e., top middle box), dark small boxes indicate components of σ_b (Eq. S6).

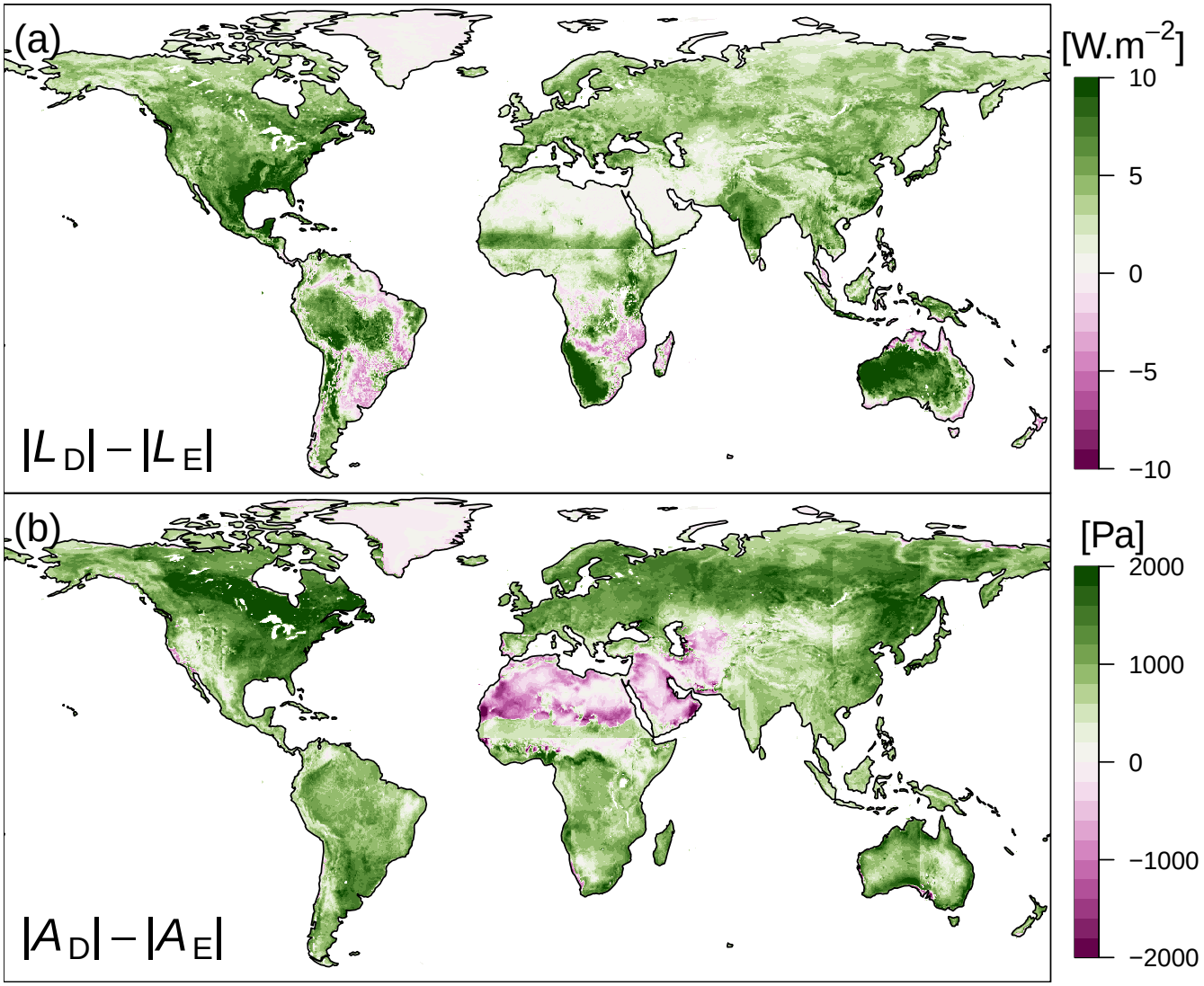

Figure S2. (a) Difference between $|\mathcal{L}_D|$ and $|\mathcal{L}_E|$ in summer (JJA and DJF for the Northern and Southern Hemisphere, respectively). (b) Same as (a) but for the atmospheric leg ($\mathcal{A}$).

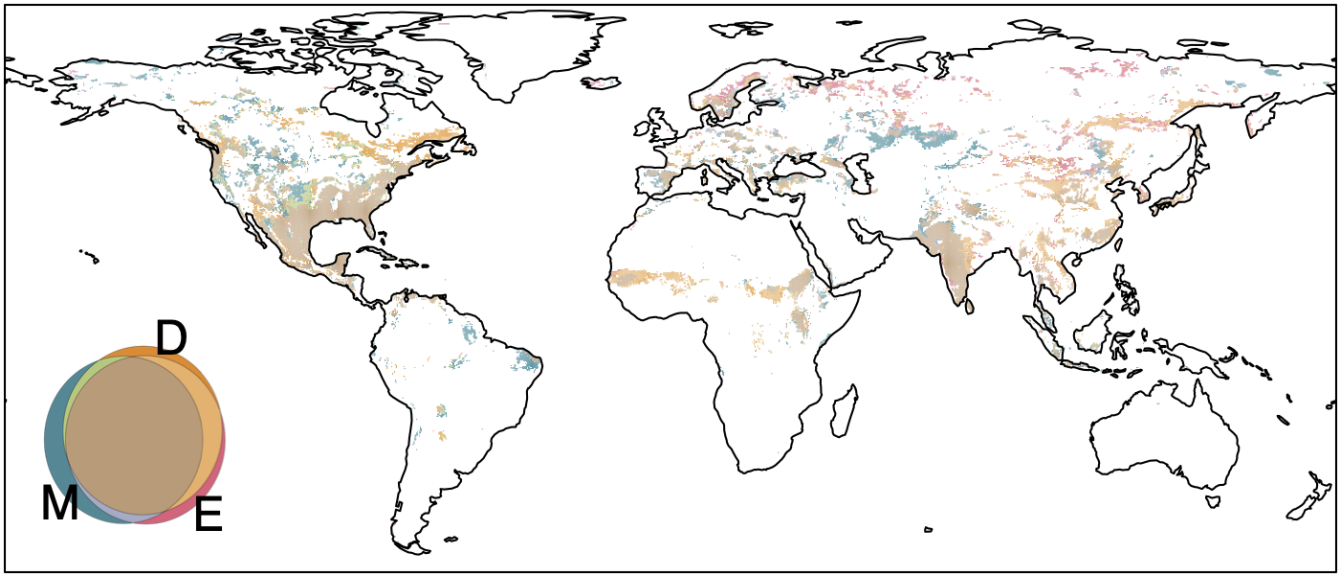

Figure S3. Spatial patterns of significant $\mathcal{L}_{\mathrm{M}}$, $\mathcal{L}_{\mathrm{E}}$, and $\mathcal{L}_{\mathrm{D}}$ (top 90% quantile of absolute values) in summer (JJA and DJF in the Northern and Southern Hemisphere, respectively). Euler diagrams show the colors for specific relationships (intersections, unions, or disjoints) among $\mathcal{L}_{\mathrm{M}}$, $\mathcal{L}_{\mathrm{E}}$, and $\mathcal{L}_{\mathrm{D}}$, and the areas of colored patterns indicate the fractions of them as well.

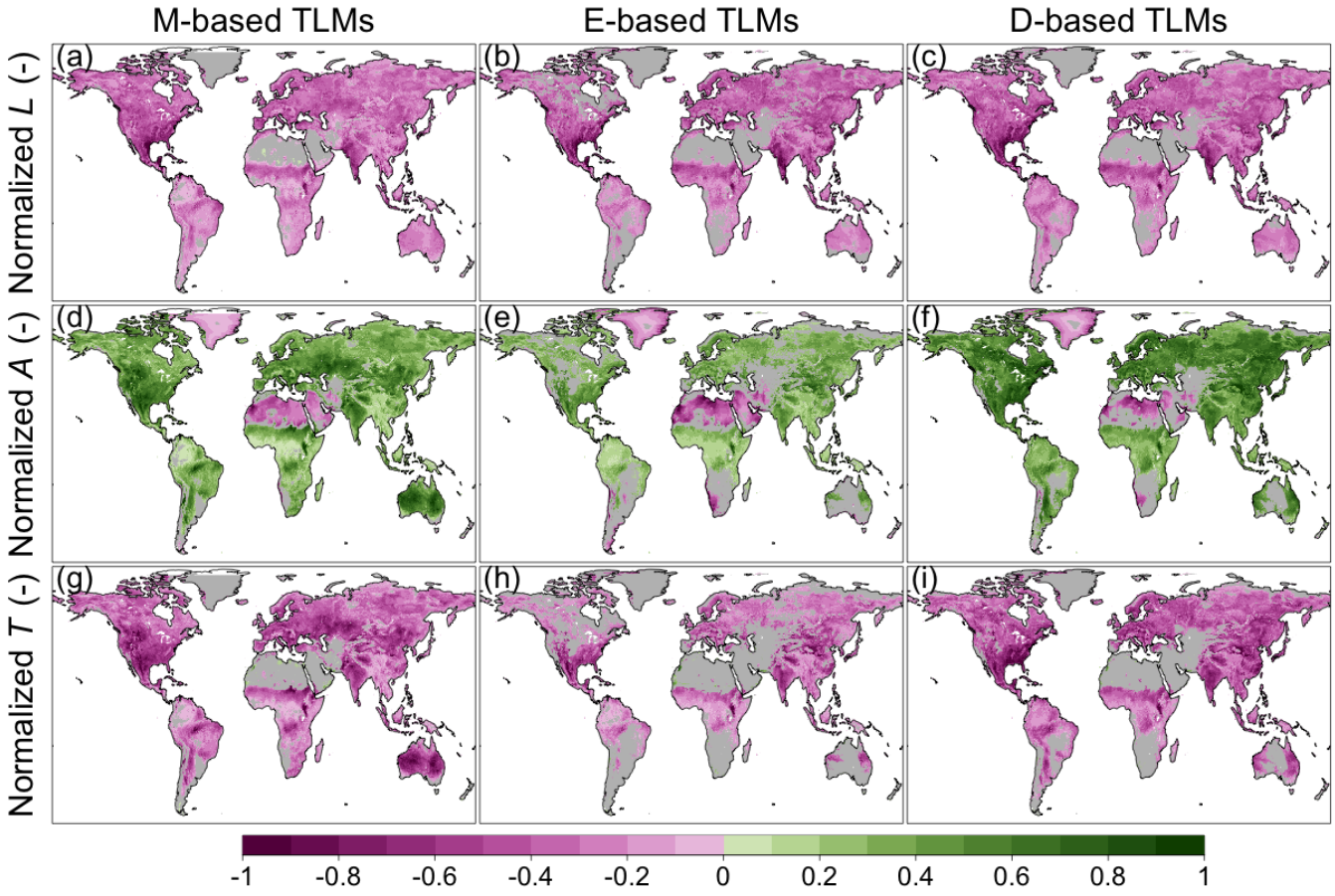

Figure S4. Maps of normalized two-legged metrics (TLMs) in JJA. Top to bottom panel: land, atmospheric, and total leg. Left to right panel: monthly-, entire-day-mean-, and daytime-only-based TLMs. To make the $\mathrm{TLM_M}$, $\mathrm{TLM_E}$ and $\mathrm{TLM_D}$ comparable, we normalize specific TLM by $n_i = \min(\max(x_i/q_{99.9\%}, -1), 1)$, where n_i indicates the normalized value of x_i and the $q_{99.9\%}$ is the 99.9% quantile of $|x_i|$. Gray regions indicate associated correlation is not significant ($p > 0.05$)

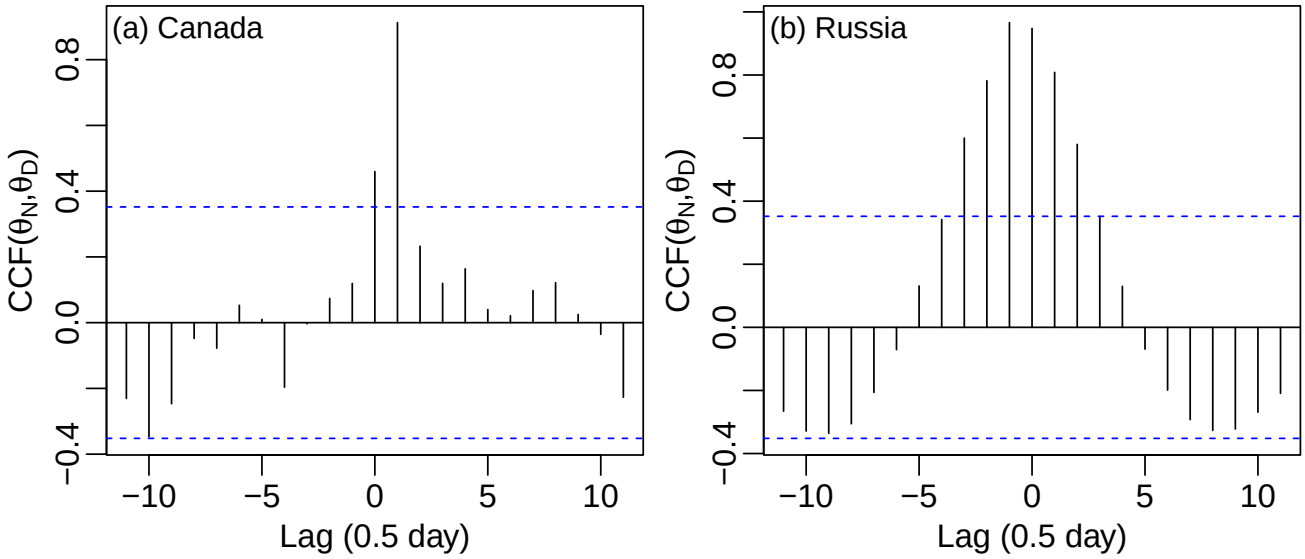

Figure S5. (a) Cross correlation function between nighttime-only-mean (N) and daytime-only-mean (D) soil moisture (θ_N and θ_D) in a grid cell located in Canada ([82.25°W, 47.5°N]). (b) Same as (a), but the grid cell is taken as a reference in Russia ([122.5°E, 68.5°N]).

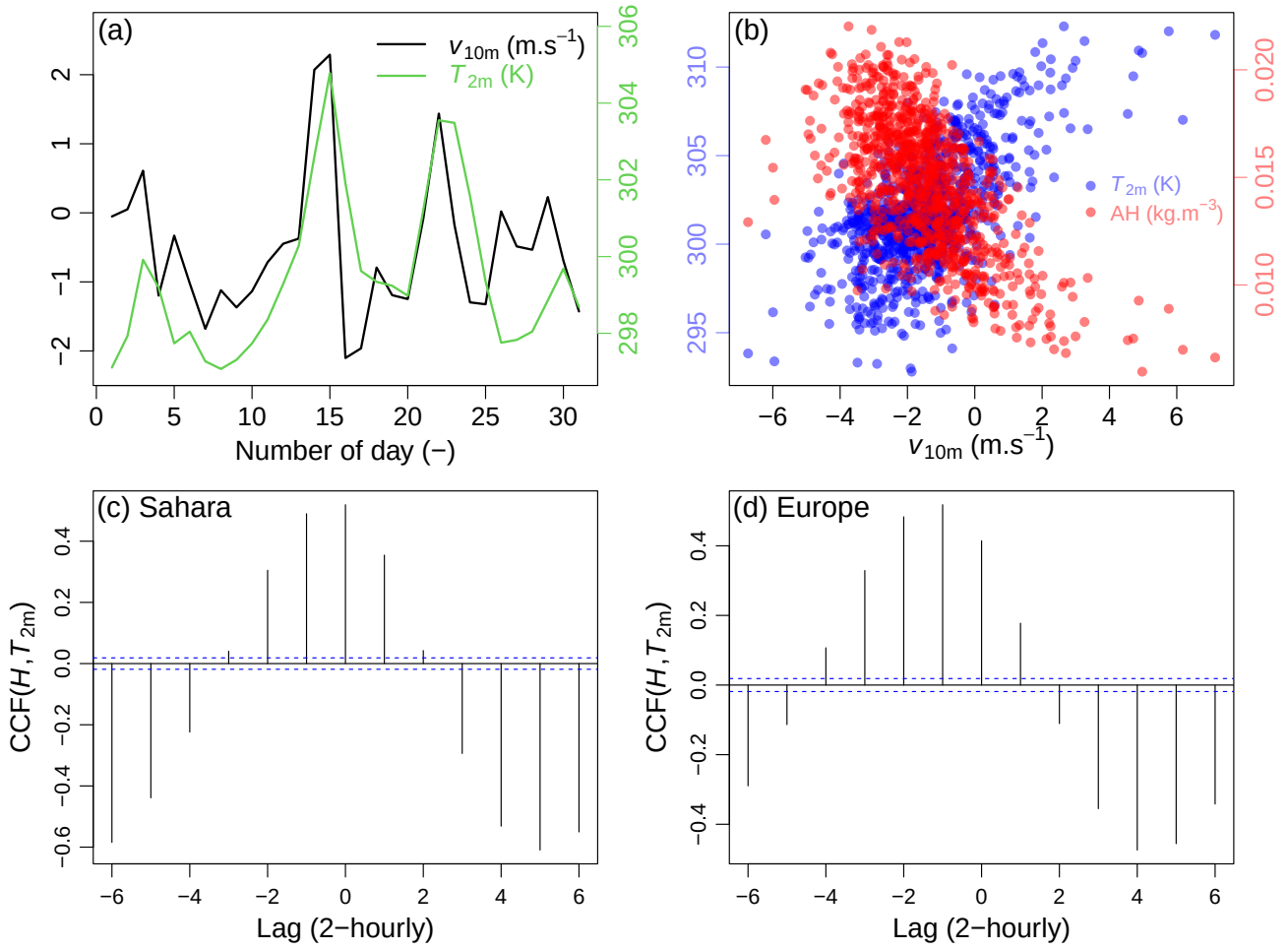

Figure S6. An example of atmospheric advection driven L-A interaction mechanism. (a) Daily 10-m northward wind speed (v_{10m}) and T_{2m} for the entire day in July 2015. (b) T_{2m} and 2m absolute humidity (AH) as a function of v_{10m}. The illustration is based on entire-day-mean daily values in JJA from 2011 to 2020. (c)–(d) Cross-covariance between two-hourly H (positive up) and T_{2m} based on two grid cells in Sahara ([12°E, 32.75°N]) and in Europe ([12°E, 47.75°N]), respectively. y-axis indicates the correlation coefficients between T_{2m} and a time-shifted H time series. The x-axis indicates the time steps of the H shifted. Negative (positive) values indicate lagged (ahead).

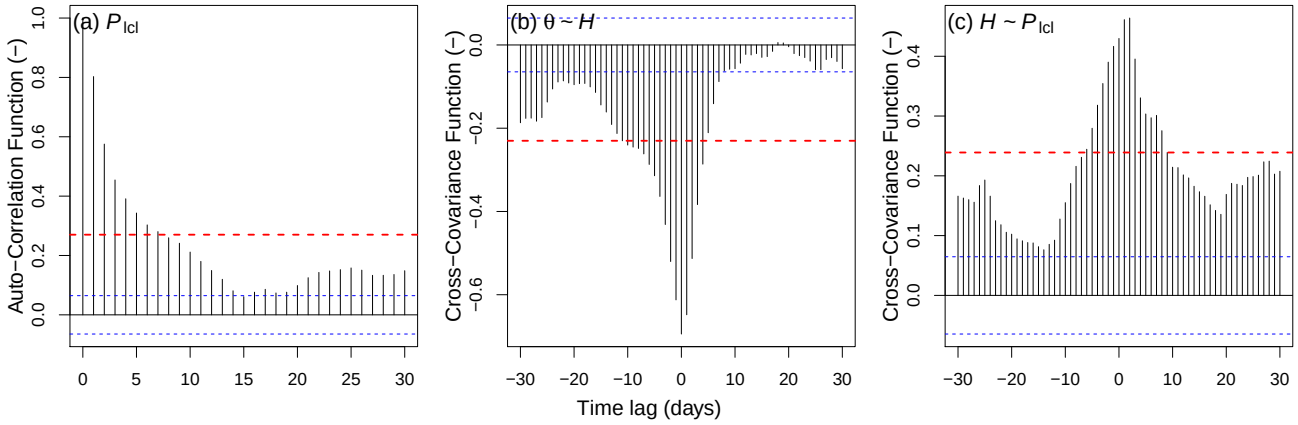

Figure S7. Examples of calculating memory indicator for the σ term and for the numerator of the ρ term ($N(\rho)$) of the two-legged metrics. (a) The entire-day-mean-based $\sigma_{P_{\mathrm{lcl}}}$ for instance, at one grid cell we first calculate the auto-correlation function (ACF) of P_{lcl} with the maximum lag of 30 days. Then the top 25% quantile of these correlation coefficients are selected (red dashed lines indicate the threshold) and averaged as the indicator $\overline{\mathrm{ACF}}_{>75\%}$. (b) For the paired θ and H, we calculate the cross-covariance function (CCF) with the maximum lag of $\pm$30 days. As the $\rho(\theta, H)$ is negative, we select the lowest 25% correlation coefficients and calculated the mean ($\overline{\mathrm{CCF}}_{<25\%}$) as indicator. (c) Similar to (b), but selecting the top 25% correlation coefficients to calculate the indicator.