[Coupling metrics to diagnose land-atmosphere interactions]

Correlation

- References:
  - Kendall, M., 1938: A new measure of rank correlation. *Biometrika*, **30**(1-2), 81–89, doi: 10.2307/2332226.
  - Pearson, K., 1895: Notes on regression and inheritance in the case of two parents, *Proc. Royal Soc. London*, **58**, 240–242, doi: 10.1098/rspl.1895.0041.
  - Spearman, C., 1907, Demonstration of formulæ for true measurement of correlation. *Amer. J. Psychol.*, **18**(2), 161–169, doi: 10.2307/1412408.
- Principle:
  - A correlation r is a measure of how one variable tends to vary (in sync, out of sync, or randomly) with another variable in space and/or time. $-1 \leq r \leq 1$.
  - The most commonly used is *Pearson product-moment correlation coefficient*, which relates how well a distribution of two quantities fits a linear regression:

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \, \sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{\overline{x^2} - \overline{x}\overline{x}}\sqrt{\overline{y^2} - \overline{y}\overline{y}}}$$

    where overbars denote averages over domains in space, time or both.
    - As with linear regressions, there is an implied assumption that the distribution of each variable is near normal. If one or both variables are not, it may be advisable to remap them into a normal distribution.
  - *Spearman's rank correlation coefficient* relates the ranked ordering of two variables in a non-parametric fashion. This can be handy as Pearson's r can be heavily influenced by outliers, just as linear regressions are. The equation is the same but the sorted ranks of x and y are used instead of the values of x and y.
  - *Kendall rank correlation coefficient* is a variant of rank correlation.
  - Under the assumption of a normal distribution, significance can be tested based on the sample size *n* (assuming independence of each value within each variable, otherwise the estimated degrees of freedom DOF < n should be used) and the value of the Student's t-test that corresponds to *n* and the desired significance level:

$$r = \frac{t_{n-2}}{\sqrt{t_{n-2}^2 + n - 2}}$$

  - The "$n-2$" is because of the 2nd-order statistics in calculating r, two DOFs are lost.
  - Usually, data are co-located in space and/or time, but this is not necessary. For example, lagged (in time) correlations can indicate delayed relationships, when a variable is lag-correlated with itself, it can indicate persistence or memory.
- Data needs:
  - Suitable for non-continuous or incomplete data as well as complete data sets.
- Caveats:
  - A strong correlation suggests a physical relationship between the two variables, although does not prove a cause-effect relationship – that can be inferred based on knowledge or hypotheses about the physical processes that relate the two variables. A correlation may also indicate an effect-effect relationship where both are responding to a common unconsidered cause.
  - If latent heat flux is not calculated in the model using a Penman-Monteith formulation, there will be a discrepancy in the diagnostics.

Supported by NASA/MAP Grant: NNX13AQ21G