THE "ATLAS OF VARIABILITY" DATA ARCHIVE

ELISABETH J. MOYER, KEVIN SCHWARZWALD, AND VICTOR ZHORIN

1. Introduction

The "Atlas of Variability" is a project by the Center for Robust Decision-Making on Climate and Energy Policy (RDCEP) at the University of Chicago to develop a computationally efficient and standardized procedure to extract the characteristics of variability in climate models. The use of spectral analysis allows for the isolation of variability in frequency-specific climate patterns. This database currently contains the output of this analysis for 12 models, 22 variables, and 3 experiment runs from the Coupled Model Intercomparison Project's 5th iteration (CMIP5). The models were chosen to be representative of the spread of model output (informed by the climate model 'genealogy' developed by [1]) and to ensure consistency in data accessibility across models.

Variability in this context is explicitly defined as the portion of the standard deviation contributed by patterns with frequencies within a defined range. Generally, this is interpreted to mean non-seasonal variability - the seasonal component is removed from the time series before processing.

The data is hosted using the Globus data publishing system. The collection is permanently stored at `https://publish.globus.org/jspui/handle/11466/102`. To access the data, you must join the RDCEP Globus Publication Users group, with instructions on how to do so to be found in the Globus interface.

1.1. Citing and Acknowledging.

To document the use and impact of this dataset, we request a specific acknowledgement of the "Atlas" in addition to acknowledgements to the authors/compilers of the original data used for this project (i.e. CMIP5 and the individual modeling groups). In text, this dataset can be referred to as the "[Atlas of Variability/AoV] [archive/database/etc...]". An example of a bibiographical entry for the entire database or large subsets thereof spanning multiple Globus 'collections' would be as follows:

Moyer, Elisabeth J., Kevin Schwarzwald, and Victor Zhorin. *The Atlas of Variability.* V0.99. July 2017. Published by the Center for Robust Decision-Making on Climate and Energy Policy (RDCEP). Supported by the NSF through the Decision Making Under Uncertainty Program. https://publish.globus.org/jspui/handle/11466/102.

For individual subsets of the collection (one model, one variable domain, etc.), the GLOBUS homepage for each dataset has a ready-made citation as a guide, i.e.

Moyer, E.J.; Schwarzwald, K.; Zhorin, V., "CMIP5 clouds variability analysis," 2016, http://hdl.handle.net/11466/RRN55BB

2. Variability Calculation

The core purpose of this data project is to calculate the frequency-separated variability of climatic variables; in other words, the portion of the standard deviation contributed to by patterns within certain frequencies. This is done through the following process (a more complete description is found in Appendix A):

(1) Detrending: a simple linear trend is removed from the time series (spectral density analysis requires stationarity)
(2) Deseasonalization: the seasonal component of variability is removed by subtracting out a 12-harmonic least-squares fit
(3) Spectral Density Calculation: a power spectrum is calculated from the resultant time series
(4) Spectral Density Integration: the standard deviation is calculated by taking the square root of the integrated spectral density between the desired frequencies

3. Data Archive Structure

The data is hosted using the Globus data publishing system. The collection is permanently stored at `https://publish.globus.org/jspui/handle/11466/102`. To access the data, you must join the RDCEP Globus Publication Users group, with instructions on how to do so to be found in the Globus interface.

To make bulk downloading easier, the data is presented in subsets across two search vectors, each containing a range of "datasets":

Models: dataset name: "[climate project] [model] variability analysis," with each dataset containing all files from a specific model (i.e. "NorESM1-M" will include files for all variables and experiments for the model NorESM1-M)

Variable domain: dataset name: "[climate project] [variable domain] variability analysis," with each dataset containing all files for a variable domain (a group of related variables; i.e. "temperature" will include files for 850 mb mean temperature and minimum, maximum, and mean near-surface temperature across all models and experiments)

Each file is therefore present in both a model subset and a variable domain subset in the Globus transfer system, and can be downloaded from either source. Model names are occasionally changed for purposes of consistency with UNIX filename conventions. For example, CSIRO-Mk3.6.0 is listed as CSIRO-Mk3-6-0.

4. File Structure

Files are named according to the following convention:

$$\begin{aligned}\texttt{filename} = &\text{[file variable identifier]\_[data frequency]\_[model]\_[experiment]\_}\\ &\qquad\qquad\text{[run]\_[start year]-[end year]\_varanalysis.nc}\end{aligned} \tag{1}$$

file variable identifier: shorthand variable name, e.g. `pr` for precipitation
data frequency: month, day, etc., using CMIP5 conventions (this includes, for example, the _Amon / _Lmon / _Omon conventions for monthly data)
model: model name in UNIX-safe character convention (so CSIRO-Mk3.6.0 becomes CSIRO-Mk3-6-0)[1]
experiment: experiment / forcing profile, e.g. RCP8.5
run: run ID (in CMIP5, [r#i#p#] for run number, iteration, and microphysics; for non-CMIP5, can be anything)
start year-end year: e.g. 2070-2099

In general, the **file variable identifier** follows CMIP5 naming practices. The biggest exception to this convention concerns variables that are 4-dimensional in the CMIP5 archive (with a height/pressure/depth level as the fourth dimension). To keep the file conventions as standardized as possible, each individual level in the 4th dimension is given its own file, and therefore its own variable identifier shorthand. For variables in which the fourth dimension is a pressure level, these are generally changed in the form `[original identifier][pressure level]` (i.e. `ta850` for 850 mb air temperature, saved as `ta` in CMIP5 parlance). Additionally, some variables are entirely constructed from CMIP5 variables, as follows:

cllow: Average cloud cover across 'low' pressure levels (using ISCCP convention,[2] > 680 mb)
clmed: Average cloud cover across 'medium' pressure levels (using ISCCP convention, 680-440 mb)
clhi: Average cloud cover across 'high' medium pressure levels (using ISCCP convention, < 440 mb)
rnett: Net top of atmosphere radiation. The positive direction is upwelling.
rnets: Net surface radiation. The positive direction is upwelling.

Each of these `_varanalysis.nc` files contains three sets of variables: global frequency-seperated variability, some basic statistics, and a geographic grid.

4.1. **Frequency-separated variability.** These variables are saved in the format `variability_[bound1]-[bound2]_[timestep]`. `[bound1]` and `[bound2]` are the period limits to the frequency band, in the units given by `[timestep]`. For example, the variability in (roughly) synoptic-scale climate patterns would be saved in a variable called `variability_3-15_days`. These variables each have three attributes - a long name (`long_name`), units (`units`), and a vector with the integration limits used when calculating the integrated spectral density (`period_bnds`). The long name attribute that contains a human-readable string description (i.e. `'3 - 15 days'` for the example above). The period

[1] Occasionally, models are listed in lowercase because their original files in the CMIP5 archive were saved as such; e.g. 'bcc-csm1-1' for BCC-CSM 1.1.

[2] Taken from the 'ISCCP Definition of Cloud Types' page by NASA, ISCCP: `https://isccp.giss.nasa.gov/cloudtypes.html`

bounds use 0 as a lowest bound by convention (though of course the lowest mathematically possible bound for patterns is 2).

4.2. Basic statistics.

In addition to frequency-separated variability, some more basic statistical properties of each pixel are saved, generally the mean and standard deviation, saved as `mean` and `std`, respectively. `std` is usually not exactly equal to the `variability_2-Inf_[timestep]` because of the limits of spectral density estimation. These variables have `long_name` and `units` attributes as above.

4.3. Geographic grid.

The geographic grid is given by 4 variables - `lat`, `lon`, `lat_bnds`, and `lon_bnds`. `lat` and `lon` identify the coordinate of each pixel, and can either be 1-dimensional vectors (rectangular grids), or 2-dimensional arrays (variable grids, etc.). `lat_bnds` and `lon_bnds` are arrays with one more dimension (of size 2) than the `lat` and `lon` variables, giving the minimum and maximum values of `lat` and `lon` that constitute the vertices of the relevant pixel. These variables have `long_name` and `units` attributes as above.

4.4. Notes on File Construction.

This dataset keeps CMIP5 naming conventions whenever possible for variable names, units, and saving conventions (including saving precipitation in kg/m^2s). For more information on variable names, units, and saving conventions, please see the files on the CMIP5 guide page here (especially the `standard_output` files). For more information on experiment design, please see the files on the CMIP5 Experiment Design Page here.

All time series used have 365-day years. For models using gregorian/leap year calendars, the 366th day of each leap year was removed before statistics were calculated.

5. Further Notes

As this method is used for further projects, we intend to expand the archive to include output from those calculations in the same format as well.

For questions/comments, please feel free to contact Kevin Schwarzwald (RDCEP, University of Chicago) at kschwarzwald@uchicago.edu.

References

[1] R. Knutti, D. Masson, and A. Gettelman, “Climate model genealogy: Generation CMIP5 and how we got there,” *Geophysical Research Letters*, vol. 40, pp. 1194–1199, Mar. 2013.

[2] W. B. Leeds, E. J. Moyer, and M. L. Stein, “Simulation of future climate under changing temporal covariance structures,” *Advances in Statistical Climatology, Meteorology and Oceanography*, vol. 1, pp. 1–14, Feb. 2015.

[3] J. M. Klavans, A. Poppick, S. Sun, and E. J. Moyer, “The influence of model resolution on temperature variability,” *Climate Dynamics*, pp. 1–11, Aug. 2016.

Appendix A. Variability Calculations

High- (i.e. the pattern of rain showers) and low-frequency (i.e. droughts, extreme precipitation events) patterns of precipitation are affected by different climatological processes, and could therefore react differently to changes in temperature and gas concentrations in addition to having differing impacts on human society. To isolate changes in variability for different precipitation processes, we use spectral density analysis to study variability in different frequency 'bands' by decomposing the time series for each model output pixel into it frequency components. To do so, we largely adapt the methodology introduced in [2] and [3] based on integrations over power spectra.

Spectral density calculations assume stationarity in the underlying time series, so we first detrend the climatic data. We isolate non-seasonal variability by several data frequency-based deseasonalization methods. Finally, we calculate band-separated variability from power spectra.

A.1. **Detrending and Deseasonalization.**

A.1.1. *Daily Data.* A classic decomposition of a time series Y is as follows:

$$\begin{aligned} Y &= X_t a + Y_c \\ Y &= m_t + s_t + Y_c \\ Y_c &= Y - m_t - s_t \end{aligned} \tag{2}$$

for trend m_t, seasonal component s_t (with a known period), and a variability component Y_c. m_t is either 0 if the time series is already stationary (as it is with equilibrated CCSM3 runs) or is made to equal 0 through subtracting a linear trend, leaving

$$Y_c = Y - s_t \tag{3}$$

The seasonal component is then removed by fitting harmonic components using least squares. Given the least-squares process

$$Y_c = Y - s_t\hat{\beta} \tag{4}$$

for

$$\hat{\beta} = \frac{s_t' s_t}{s_t' Y}$$

we fit

$$s_t = \left[1 \cos\left(\frac{2\pi}{365} \begin{bmatrix} 1 \\ \vdots \\ t \end{bmatrix} [1 \ldots \lambda] \right) \sin\left(\frac{2\pi}{365} \begin{bmatrix} 1 \\ \vdots \\ t \end{bmatrix} [1 \ldots \lambda] \right) \right] \tag{5}$$

for length of (daily) time series t and number of harmonics to be removed λ. In the general process for daily time series used in this project, $\lambda = 12$, and for 30-year time series, $t = 10950$. Subtracting $s_t\hat{\beta}$ from equation 4 above results in Y_c now representing the detrended, deseasonalized, variable component of the time series, ready to be further analyzed.

A.1.2. *Monthly Data.* Monthly data was deseasonalized by taken the simple average of each month over the length of each time series and subtracting it from every data month. Y_c, the deseasonlized time series Y over T years, was constructed as follows:

$$Y_c(y,m) = Y(y,m) - \frac{1}{T}\sum_{y=1}^{T} Y(y,m) \tag{6}$$

for each month m and year y.

A.2. **Spectral Analysis.** The autocorrelation function for a stationary process $x(t)$ with mean $\mu = 0$ and variance σ^2 is given by

$$R(\tau) = \frac{1}{\sigma^2}E[(x_t - \mu)(x_{t+\tau} - \mu)]$$
$$R(\tau) = \frac{1}{\sigma^2}E[x_t x_{t+\tau}]$$

and is periodic at the same period as the original function $x(t)$. Peaks in $R(\tau)$ correspond to periodicites with frequencies τ - the autocorrelation function finds interior periodicites in the original time series. By the Wiener-Khinchin Theorem, the autocorrelation function makes a Fourier Pair with the power spectral density $S_{xx}(\omega)$ as follows:

$$S_{xx}(\omega) = \int_{-\infty}^{\infty} R(\tau)e^{-i\omega\tau}d\tau$$

The absolute value of the Fourier Transform as a function of frequency gives the amount of that frequency that is present in the original function, in this case $R(\tau)$. Therefore the (infinite) sum of the Fourier transforms over a range of frequencies gives the contribution of those frequencies to the autocorrelation, which gives how strongly different frequency patterns show up in the time series. In other words,

$$\begin{aligned} \int_{\omega_1}^{\omega_2} S_{xx}(\omega)d\omega &= \frac{1}{T}\int_{\omega_1}^{\omega_2} |\hat{x}(\omega)|^2 d\omega \\ &= \text{contribution to power by } \omega \in [\omega_1, \omega_2] \end{aligned} \tag{7}$$

Now, taking the sample variance of a discrete stationary time series x of length N with mean $\mu = 0$,

$$\sigma^2 \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i)^2$$

and assuming an infinite time series ($N \to \infty$), we see that the variance of a time series is related to its average power $\bar{P}$ over the domain $[-T, T]$ through

$$\sigma^2 = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(x_i)^2 \to \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} x(t)^2 dt = \bar{P}$$

Using Parceval's theorem, which states

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{x}(\omega)|^2 d\omega$$

for the Fourier transform $\hat{x}(f)$ of $x(t)$, we can see that

$$\frac{1}{2T} \int_{-\infty}^{\infty} |\hat{x}(\omega)|^2 d\omega = \frac{1}{T} \int_{0}^{\infty} |\hat{x}(\omega)|^2 d\omega = \sigma^2$$

Combining this expression with equation 7 above,

$$\frac{1}{T} \int_{\omega_1}^{\omega_2} |\hat{x}(\omega)|^2 d\omega = \sigma^2 \{\omega \in [\omega_1, \omega_2]\} \tag{8}$$

with the standard deviation contained in those frequencies simply given by the square root of the expression,

$$\sigma(\vec{\omega}) \equiv \sigma\{\omega \in [\omega_1, \omega_2]\} = \sqrt{\frac{1}{T} \int_{\omega_1}^{\omega_2} |\hat{x}(\omega)|^2 d\omega} \tag{9}$$

Center for Robust Decision-Making on Climate and Energy Policy (RDCEP)