Not an Average Visualisation Project

J. W. Larson[1]   P. R. Briggs[2]

[1]Mathematical Sciences Institute, The Australian National University

[2]CSIRO Marine and Atmospheric Research

29 October 2014

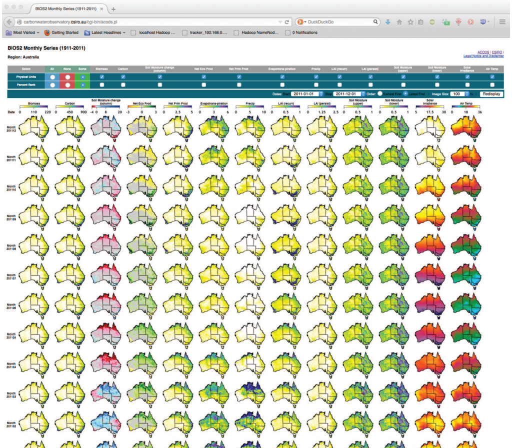
during the next frantic twenty minutes...

outline

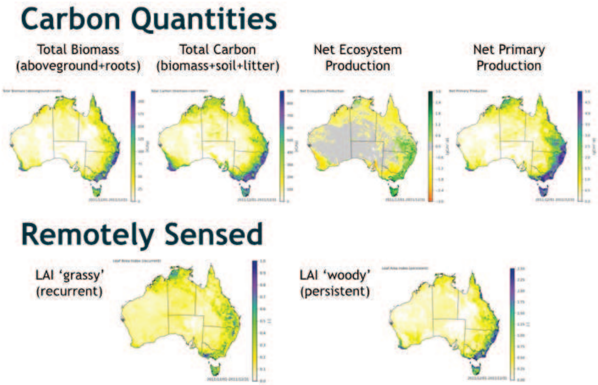
- Australian Carbon Water Observatory—The kind of dataset we should be developing
  - Surface Carbon and Water budgets
  - Relationship to Australian Water Availability Project (AWAP)
- Building the elements of the Observatory–just the visualisation part
  - File and data handling, a.k.a. "Data Janitor Work"
  - Production of literally millions of images using high performance computing.
  - Relationship to Another Big Problem: Data Science
- Mining variability information from ACWO data
  - Background: Time-dependent Probability Density Functions
- Summary

australian carbon water observatory

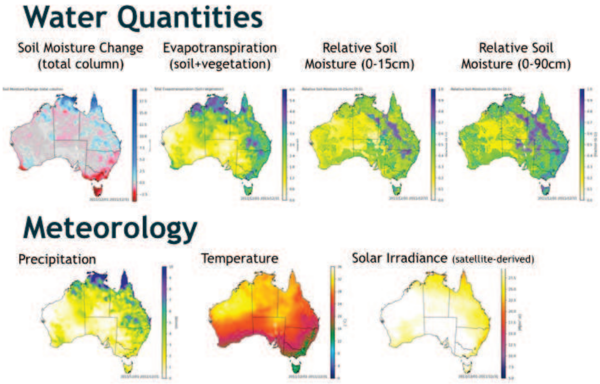
australian carbon water observatory (national scale)



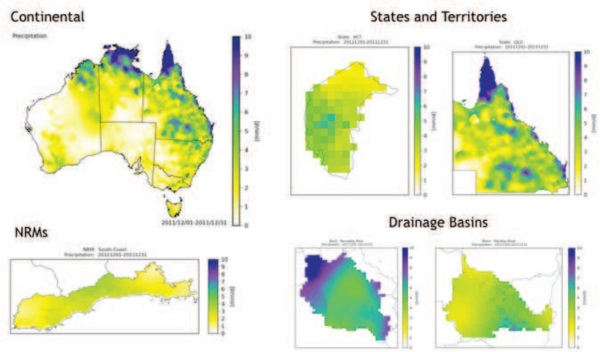
australian carbon water observatory—carbon



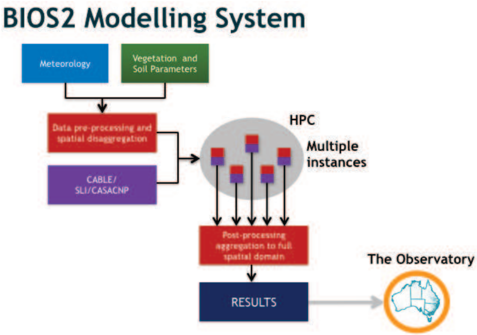
australian carbon water observatory—
water and meteorology



australian carbon water observatory—
...at multiple scales...



australian carbon water observatory—how



data janitor by day...

superhero by night...

- Massive visualisation project utilising Python's rich ecosystem
- Run on HPC platforms with MPI parallelism
- Data format, grid, file naming conventions compatible with AWAP
- There's a lot more we can do with these

HOME SEARCH The New York Times SUBSCRIBE NOW

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Monica Rogati, Jawbone's vice president for data science, with Brian Wilt, a senior data scientist. Peter DaSilva for The New York Times

EMAIL

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science

time-dependent probability density functions (tdpdfs)

what is it, and how do I make one?

- **Given:** A meteorological timeseries $X(t)$ and some sampling interval $[t_0, t_0 + W)$
  - Sample spans a total of Y years
  - W is the *sampling window width*—assume a square window
  - Quantities t_0, W measured in years; Pick W by convention
- Slide a window of width $W = 30$ years through the data, displacing it one year at a time
- Estimate the climate PDF for every unique 30-year window, assigning the year of the window's center as the "time" for that climate PDF
- **Result:** $\rho(X, t)$ is a collection of $Y - 30$ "time slice" PDFs
  - $\rho(X, t) \geq 0, \quad \forall\ X, t$
  - For fixed $t = t_p$, $\int_{-\infty}^{\infty} \rho(X, t_p) dX = 1$.

pdfs and information theory

shannon entropy H

$$H(X) = \int_{-\infty}^{\infty} \rho(X) \log \rho(X) dX$$

- $H(X)$, quantifies the amount of "surprise" present in X.
- Logarithm base defines units—*bits* for base 2, *nats* for base e

kullback-leibler divergence $D_{KL}(\rho||\psi)$

Given a second density function $\psi(X)$ that models $\rho(X)$,

$$D_{KL}(\rho||\psi) = \int_{-\infty}^{\infty} \rho(X) \log \left(\frac{\rho(X)}{\psi(X)} \right) dX$$

- Also called the *KL Gain*, because it's how much additional information, given $\psi(X)$, is required to describe $\rho(X)$

tdpdfs, information theory, and variability

applications

- Time-dependent probability density functions
  - At-a-glance time history of the density, potentially revealling interannual-scale to interdecadal-scale structure
- Kullback-Leibler Divergences
  - Marginal informativeness of extending climate sampling intervals; i.e. adding a year to a sampling window
  - (Un-) Representativeness of a subset of the climate record when used to model the full record
  - (Un-) Representativeness of the full climate record when used to model a subset of it
  - Interdecadal variability /non-stationarity of the density function

case study: central england temperatures

the longest weather station-based observational record

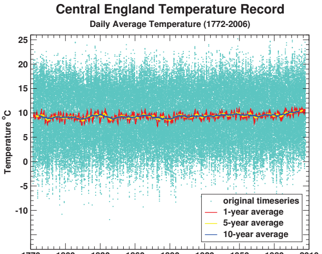
Surface air temperature computed from three stations representative of Central England

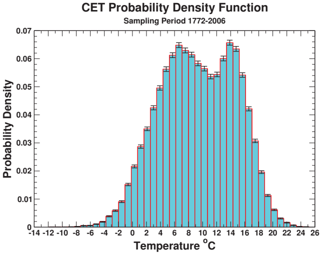
- Manley's Monthly Average record runs 1659–present, Parker et al. have computed Daily Averages (1772–present) and Extrema (1878–present)
  - Precision of 0.1°C for 1722–present, 0.5°C before that
- The daily and monthly CET data are known to have these properties:
  - Oscillatory behavior on multiple scales up to century and beyond
  - Long-term warming trend
  - Also urbanization-related warming, but this is corrected

central england temperatures

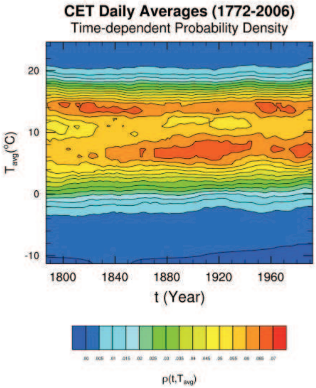
pdf estimation method

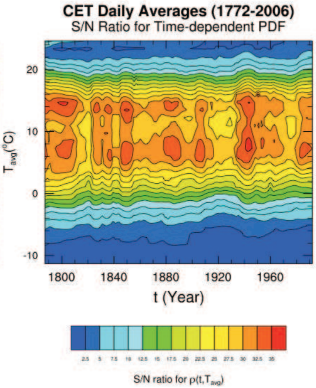
- Piecewise-cosntant PDF estimated using Bayesian-based optimal binning (Knuth, 2005)
  - Produces a number of uniform bins that is the most honest reflection of the sample



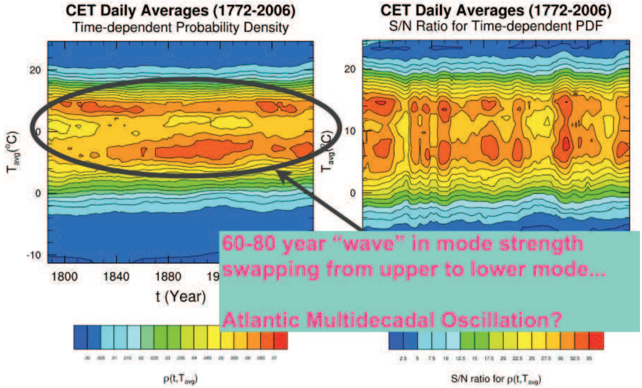


time-dependent pdf for cet daily T_{AVG}

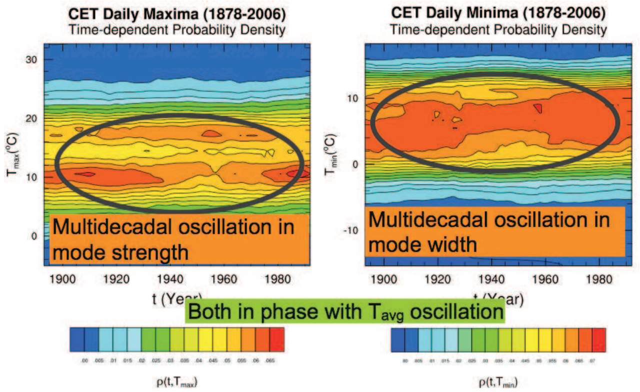




time-dependent pdf for cet daily T_{AVG}



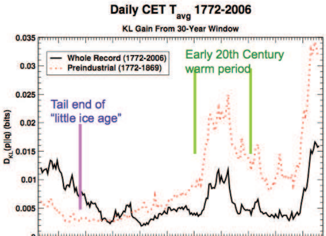
time-dependent pdf for cet daily T_{MIN} and T_{MAX}

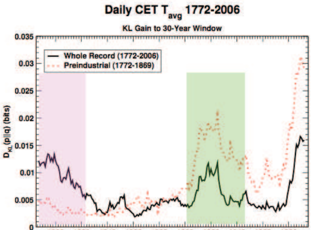


cet—representativeness of long or window records (or lack thereof)

the kld quantifies differences between pdfs

- Comparison of 30-windowed PDFs with
  - The Preindustrial Era
  - The Full daily CET Record
- The lower the value of the KLD, the better the agreement

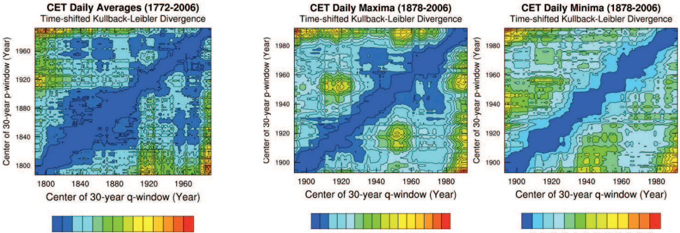




cet—long-term variability

the kld quantifies differences between 30-year window pdfs

- Plots show the relative lack of skill for a PDF constructed from a window centered on the ordinate value for modelling the PDF constructed from a sample centered on the the abcissa value
- The lower the value of the KLD, the better the agreement

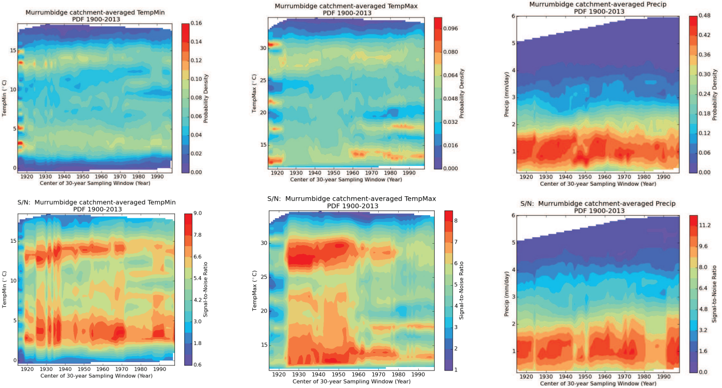


case study: multivariate variability estimation for awap

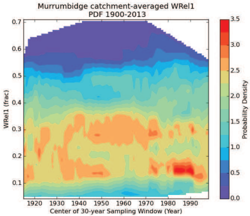
sample data

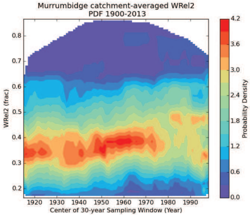
- Data taken from AWAP Historical Run 26j
- Timeseries of catchment area-averaged variables for the Murrumbidgee Catchment (1900-2014)
  - Meteorological drivers: minimum/maximum temperature and precipitation—(T_{min}, T_{max}, Pr)
  - AWAP outputs
    - Upper (W_{Rel1}: 0–0.2m) and lower (W_{Rel2}: 0.2–1.5 m) layer soil moisture
    - Total evapotranspiration: $F_{WE} = F_{WTra} + F_{Wsoil}$
    - Total discharge: $F_{WDis} = F_{WRun} + F_{WLch2}$

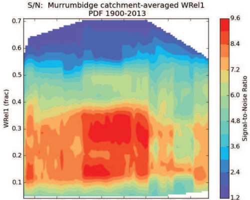
awap 26j tdpdfs—meteorological drivers

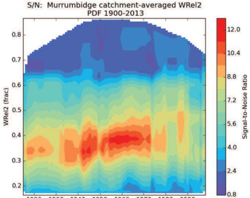


awap 26j tdpdfs—soil moisture

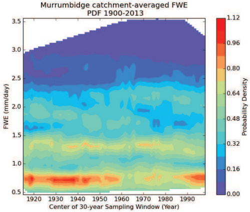


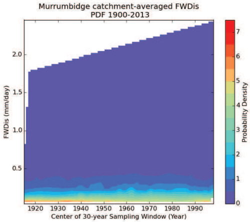


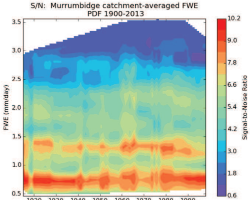


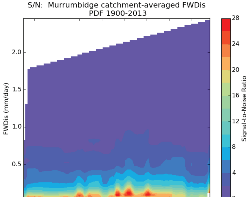


awap 26j tdpdfs—evapotranspiration and drainage

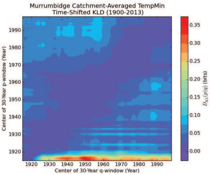


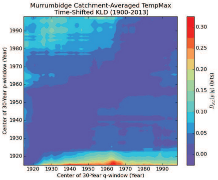


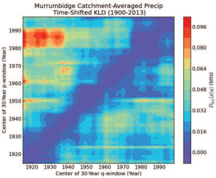




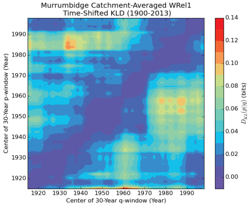
awap 26j long-term variability—meteorological drivers

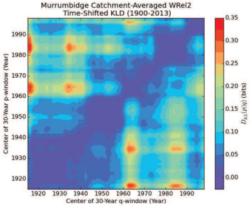


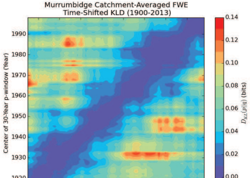


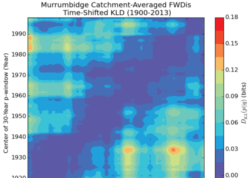


awap 26j long-term variability—evapotranspiration and drainage









discussion

kind of a mixed bag

- The technique does reveal some interesting things
    - Picks up the 1900-1910 climatology in the meteorological drivers
- Disappointments include
    - Low signal-to-noise ratios in the time-dependent PDFs
    - Ambiguity in signals / absence of pronounced structure in the time-dependent PDFs
- Much of the disappointment is likely a function of catchment-scale averaging of the timeseries.

summary

conclusions

- We have built a carbon water observatory for Australia with rich graphical content
  - Millions of monthly average images across continental, state, catchment and natural resource management regional scales
- We have presented a preliminary example of the kind of deep, multivariate, spatiotemporal analysis that is now possible

future work

- Complete rollout of images and animations
- Make data and open-source software tools available
- Pursue this relatively shallow "deep" analysis more deeply
  - Gridpoint timeseries sampling
  - Spatial maps of interdecadal Kullback-Leibler Divergences
  - Uncertainty quantification