The climate data analysis challenge

Working Paper/Methodological guidance note

Dr Chris Jack, University of Cape Town

The urban risk and decision making context creates a unique challenge to climate information analysis, production, and communication. Urban climate or weather related risks are complex and typically emerge through poorly understood interactions across spatial and temporal scales as well as sectoral contexts. For example, health issues can emerge as a result of increased dependence on shallow wells contaminated by human waste in urban and peri-urban residential areas resulting from failed formal water supply driven a by a complex combination of infrastructural failure, management failures, remote catchment rainfall failure driving hydro-power failure, impacting deep groundwater pumping capacity.

Providing climate information in such a context requires multiple types of climate data from multiple sources, suitably analysed and processed for quality and statistically relevant signals. While much work is focussed on co-production and multi/trans-disciplinary modes of climate information or knowledge generation, this work often rests on fundamental data access and analysis which in itself is a growing challenge.

The ability to rapidly mine large climate datasets as multi-disciplinary processes evolve in a particular context is becoming increasing critical. It is a significant obstacle to multi-disciplinary processes when new climate information cannot be generated timeously. The absence of this capacity results in a strong dependence on pre-generated climate information generated within highly resourced, typically northern institutions, strongly decoupled from real world applications.

An example of this is the proliferation of climate information portals or websites providing a diversity of contradictory climate data with commonly little guidance regarding suitability to different applications or geographical areas[1].

Usable but powerful climate analysis software is one component in changing this picture. Such software can enable local climate researchers to easily generate more appropriate climate information within their own contexts rather than relying on externally pre-generated

[1] Hewitson, B.C.,Waagsaether, K.L., Wohland, J., Kloppers, K., Kara,T., The evolving landscape of Climate Information Websites, WIREs, under review. 2016

data. However, the use of such software does require greater technical skill and so may need to be supported through capacity development activities. However such capacity development is an added opportunity to increase local capacity to not only analyse data but also interpret and communicate the results of the analyses.

Climate data challenges

There are a number of key challenges related to mining climate data for research or decision relevant information:

- Volume: The development and explosion of satellite based observations datasets and derivatives, as well as the exponential increase in the volume of climate model data has resulted in unprecedented volumes of data that exceed the typical storage capacity of all but the most highly resourced institutions, and arguably have exceeded the analysis capacity of climate science generally.[2]
- Complexity and diversity: The proliferation of data has been paralleled by an increase in complexity and diversity of data and data production methodologies. Observations are now derived through complex algorithms and the resultant data often not intuitive to use[3]. New data storage formats have been developed to respond to large data size but these formats often make using the data challenging.
- Contradictions: With diversity comes contradictions. Across Africa roughly 16 different observed rainfall products are available but analysis of trends across these products results in a wide range of messages. These contradictions are often not forefronted in analysis, not well understood, and not well communicated.

The role of data analysis software

Solutions to the challenged detailed above are not simple or uniform across contexts. However, analysis software has a role to play in addressing some of the challenges. It is also clear that as data volumes and diversity grows, analysis software needs to evolve appropriately. It is also clear that software cannot replace sound climate science understanding. Software is merely a tool to support climate information production.

An example of analysis software that has played a significant role in climate analysis, specifically analysis of historical extremes, is the ETCCDI software. This code, written in the R scripting language allows for the statistical analysis of observed time series in order to identify key statistics such as temperature and rainfall extremes, diurnal temperature range, and maximum consecutive dry spells. The ETCCDI indices are comprehensive and have become a defacto standard suite of extremes and more general climate statistics. However, arguably the real power of the indices was largely realised through the distribution of the analysis software which has allowed a diverse community of researchers and practitioners to apply the analysis on their own data and within their own analysis frameworks rather than rely on a suite of pre generated data.

[2] The Coupled Model Intercomparison Project (CMIP) version 3 produced around 36TB of data. CMIP5 produced around 2PB (2,000 TB) of data. CMIP6 may produce 500PB (500,000 TB)

[3] A recent review of observed data for Africa identified around 16 different gridded observed rainfall products.

The ETCCDI software only supports station time series data and so has limited application. While many researchers re-implement the statistics for different applications this takes time and resources which are not always available.

Another more recent example of such analysis code is the Open Climate Workbench originated out NASA JPL and focussed on facilitating the analysis of the CORDEX downscaled climate model data.

Similarly, the Open Climate GIS (OCGIS) package enables GIS style analysis of climate datasets in response to the growing use of GIS modeling to inform decision making and the need to integrate climate data into GIS modeling, a task that has traditionally been quite challenging.

The *climstats* analysis code is being developed within the UrbanARK project. Initial development has been focussed on three particular challenges:

- Analysis of diverse observed datasets, both gridded and point time series in multiple data formats including new high performance distributed storage architectures.
- Flexible vulnerability relevant statistics on multiple time scales
- Area aggregation across complex spatial domains.
- Dealing with complexities of model data including non-standard calendars

Unlike the ETCCDI software, *climstats* does not produce a list of standard statistics but rather allows the user to perform a smaller number of statistical analyses with user provided thresholds and/or parameters. So, whereas the ETCCDI indices include an analysis of the number of days per year warmer than 25C as a “summer days” index, *climstats* allows the user to specify arbitrary thresholds relevant to the analysis context. In many contexts in Africa, the number of days exceeding a threshold such as 35C might well be far more relevant.

Visualisation

Soon to be included in climstats are a number of new visualisation methods. Visualisation has been shown to be a particularly important step providing information to decision makers. One in particular new visualisation method soon to be included in climstats, is the plume plot.

An example plume plot can be seen below (figure 1) showing projections of the number of days per season exceeding 35C in Ibadan, extending from the recent past through to the end of the 21st century. The plume plots are especially useful in describing multi-model spread but also emergence of climate change signal from background natural variability. The color shifts from blue to green are used to indicate when a particular model (the individual lines) begins to exhibit a climate that is unexpected relative to recent past.

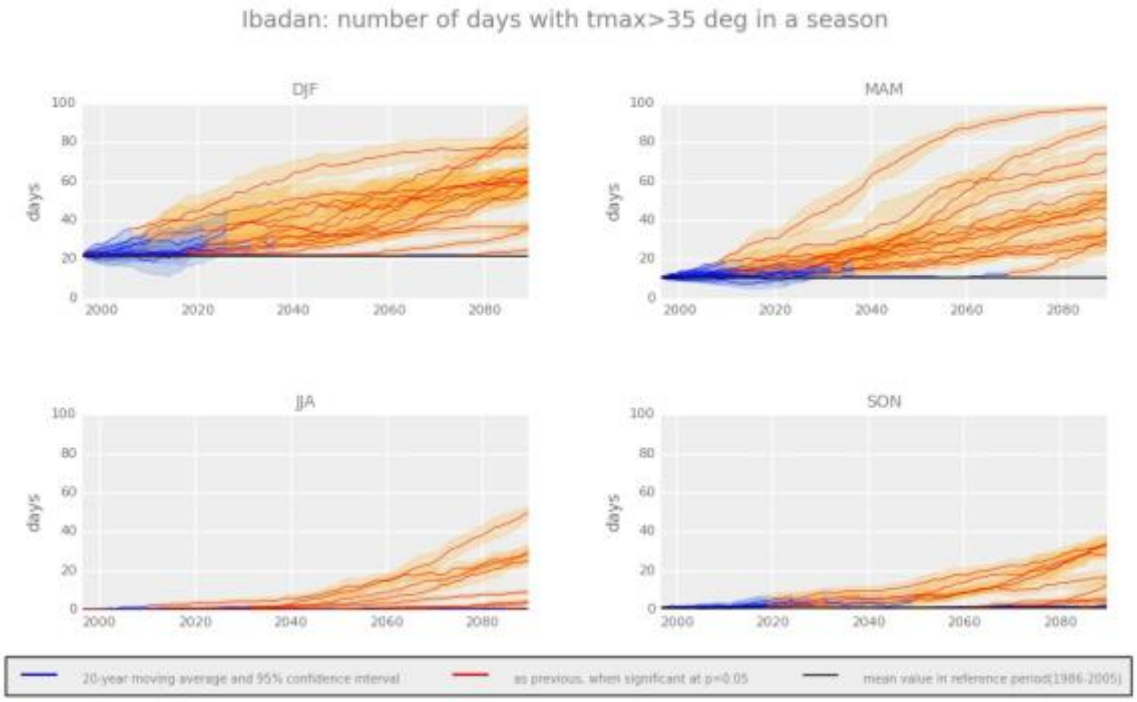

Figure 1: Downscaled projected changes in the number of days per three month season exceeding 35C in Ibadan, Nigeria.

Typically temperature changes emerge clearly from natural variability far sooner than rainfall changes. The figure below (figure 2) is similar to the one above in that it shows downscaled projections for Ibadan, but in this case the statistic is the number of days per season with heavy rainfall (top 10th percentile of daily magnitudes). It can be seen that this far more disagreement between models but also that even for many individual models, very little noticeable change is seen right through the century.

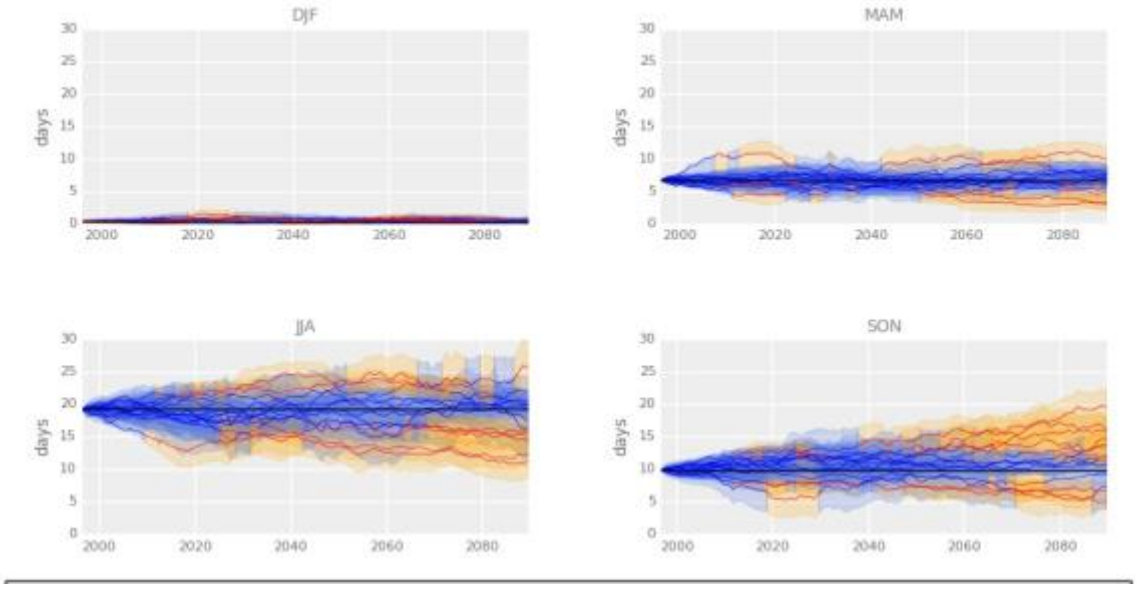

Figure 2: Downscaled projected changes in the number of days per three month season exceeding the 90th percentile of daily rainfall amounts in Ibadan, Nigeria.

Data format support

A core component of the *climstats* code is the data access support layer. This layer of code has been explicitly designed to support multiple data storage formats including the easy addition of new formats as demand arises or new formats are developed. Currently NetCDF and ASCII data storage formats are supported but support for high performance distributed data storage such as SciDB[4] and Rasdaman[5] are being experimented with in order support high performance distributed computing architectures.

Another avenue of interest is Web Processing Services (WPS)[6] and related systems that allow data processing to be "off loaded" to large centralised computing facilities. Such systems are now being developed within the Earth System Grid Federation (ESGF) architecture that supports the CMIP5 and soon the CMIP6 climate projections archives. This is a very exciting possibility for under resourced researchers as it gives them the flexibility to develop their own analysis, while allowing those analyses to run on large centralised computing facilities possibly even in a different country or continent.

The *climstats* code has been explicitly developed with a view to supporting such new and innovative solutions to the growing challenges of climate data analysis.

Implementation

The *climstats* code is implemented in the very common and powerful Python programming language. Python is widely supported across all computing platforms (Microsoft, Apple, Linux) and is rapidly becoming the most common scientific data analysis languages. The core language is now enhanced through a huge library of supporting code packages providing statistics, visualisation, database, mapping, GIS, and many other types of functionality.

Currently the implementation is in the form of a Command Line Interface (CLI). While less user friendly, CLI have the distinct advantage of being simple and well defined as well as supporting batch processing systems that allow for large analysis tasks that require many hours to run, to be configured.

As *climstats* is developed further two avenues will be explored. The first is to implement a web service interface similar to the current CSAG Climate Information Platform (CIP) but more directly focussed on user analysis of climate data, the second is to tenatively explore the development of a Graphical User Interface (GUI) for *climstats* to make it more accessible to non-technical users. However, currently the code development is focussed on the data analysis and visualisation component.

[4] SciDB: http://www.paradigm4.com/about/

[5] Rasdaman: http://www.rasdaman.org/

[6] https://en.wikipedia.org/wiki/Web_Processing_Service