Environmental Modelling & Software 96 (2017) 181–198

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A toolkit for climate change analysis and pattern recognition for extreme weather conditions – Case study: California-Baja California Peninsula

Saeid Ashraf Vaghefi*, Nazanin Abbaspour, Bahareh Kamali, Karim C. Abbaspour

Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Duebendorf, Switzerland

A R T I C L E I N F O

Article history:
Received 5 July 2016
Received in revised form
14 June 2017
Accepted 25 June 2017

Keywords:
Big data
Customized data mining
Bias correction and statistical downscaling
Spatial interpolation
Critical consecutive days analyzer
Extreme events
ISI-MIP

A B S T R A C T

This paper describes the development of a Climate Change Toolkit (CCT) to perform tasks needed in a climate change study plus projection of extreme weather conditions by analyzing historical weather patterns. CCT consists of Data Extraction, Global Climate Data Management, Bias Correction and Statistical Downscaling, Spatial Interpolation, and Critical Consecutive Day Analyzer (CCDA). CCDA uses a customized data mining approach to recognize spatial and temporal patterns of extreme events. CCT is linked to an archive of 0.5° historical global daily dataset (CRU, 1970–2005), and GCM data (1960–2099) for five models and four carbon scenarios. Application of CCT in California using ensemble results of scenario RCP8.5 showed a probable increase in the frequency of dry periods in the southern part of the region, while decreasing in the north. The frequency of wet periods may suggest higher risks of flooding in the north and coastal strips. We further found that every county in northern California may experience flooding conditions of 1986 at least once between 2020 and 2050.

Software availability

Software Name: Climate Change Toolkit (CCT) (version 1.1.1)
Developers: Saeid Ashraf Vaghefi, Karim Abbaspour
Database: Historical (1970–2005) daily CRU TS 3.1 precipitation and temperature data, global 0.5° grid, ASCII format
Database: 5 GCMs (CMIP5), 4 Scenarios (1960–2099), daily precipitation and temperature data, global 0.5° grid, ASCII format
Year first available: 2017
Software required: Microsoft Excel and Access
Programming Language: C#
Operating system: Windows XP or newer
Availability: Software and data can be downloaded from www.2w2e.com. Ask authors (Saeid Ashraf Vaghefi and Karim Abbaspour) for codes and contribution for further developments
License: GNU General Public License
Cost: Free

1. Introduction

Climate change is now widely recognized as one of the major 21st century environmental problems facing the globe. Atmosphere Ocean General Circulation Models (GCMs) suggest that rising concentrations of greenhouse gases will have significant implications for climate worldwide (IPCC, 2007). Assessment of the climate change impact and mitigation at both global and regional scales has attracted the attention of decision makers, researchers, and stakeholders (Yohe et al., 2007). Impact assessment requires climate data at various spatial and temporal scales (IPCC, 2013). Although GCMs are typical sources of future climate data, at the moment, their spatial resolutions are often too coarse for regional impact studies (IPCC, 2013; Wilby et al., 2002). Moreover, all GCM outputs involve large biases that, if not corrected, can lead to significant errors in regional hydrological impact assessments (Teutschbein and Seibert, 2012). Handling and processing of big databases of high spatial resolution and extended time series, bias correction and downscaling of GCMs outputs, and spatial interpolation of climate data to achieve finer resolution are typical tasks in almost every climate change study (Ahmed et al., 2013). These tasks make climate change studies time consuming, repetitive, and tedious (van Vuuren and Carter, 2014).

* Corresponding author. Ueberlandstr 133, 8600 Dubendorf, Switzerland.
E-mail address: saeedashrafv@gmail.com (S. Ashraf Vaghefi).

http://dx.doi.org/10.1016/j.envsoft.2017.06.033
1364-8152/

Climate change has been shown to have significant impacts on hydrological processes worldwide (Abbaspour et al., 2009; IPCC, 2013; Krol and Bronstert, 2007; Krysanova et al., 2015; Reyenga et al., 1999; Risbey, 2011; Teegavarapu, 2010; Vaghefi et al., 2014; Vorosmarty et al., 2000). Lehner et al. (2006) analyzed the effect of climate change on the frequency and severity of extreme hydrological events such as droughts and floods. Extreme hydrological events have devastating impacts on human lives and the environment. Expected increases in extreme weather increase the risk of death and destruction (IPCC, 2007). From May to September of 1999–2003, Basu and Ostro (2008) reported that for every 5.6 °C increase in temperature, there was a 2.6% increase in cardiovascular deaths in 9 California counties. The frequency of extreme events across the world is changing. Recent years have seen a larger number of severe droughts and floods in different regions around the world, resulting in depletion of water resources, and adverse impacts on agricultural and economic sectors (Berg and Hall, 2015; Lehner et al., 2006; Peterson et al., 2002).

In this study, we describe the Climate Change Toolkit (CCT), which was developed with several objectives in mind. These were *i*) handling of big data, as it is required by climate change analyses, especially at large scales and long time periods, *ii*) easy and seamless calculation of necessary steps in climate change studies, such as data reformatting, data interpolation, downscaling and bias correction, and *iii*) projection of historical extreme events into the future by pattern recognition of past events. These objectives are elaborated in more detail below.

i) Big data is defined as any collection of data sets where volume and complexity make processing difficult using traditional tools (Vitolo et al., 2015), hence requiring advanced methods to extract values from databases (Hilbert and Lopez, 2011). Big data require infrastructure for storing and managing to facilitate further analyses and post processing.

The so-called 4 V's of big data is defined by IBM as *Volume*, *Variety*, *Velocity* and *Veracity* (http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data). *Volume* is the large amounts of data becoming available through new technologies (Lokers et al., 2016), *Velocity* is the high processing time required to make the data available for end-users and decision makers (Lokers et al., 2016). *Variety* is the varied type of data relevant for decision making (Lokers et al., 2016), and finally, *Veracity* is the integrity and accuracy of data and their sources. Lokers et al. (2016) reviewed the previous studies in the context of big data in the agro-environmental field and concluded that there is a vital need for frameworks and working procedures that ensure the integrity of data and its derived added-value products. To frame the way big data is used in decision, Lokers et al. (2016) introduced a knowledge management model called "Data-information-knowledge-wisdom" or "DIKW hierarchy" to conceptualize the process of turning the massive data, which as a raw material has little or no significance to end users, into compact, structured and contextualized, manageable blocks that are applicable in a specific decision making context.

The outputs of GCMs are time series of climate data such as daily precipitation, and maximum and minimum temperature, which are usually extended, high frequency data and need preprocessing and some tools for summarizing before further analyses. These tasks are in the context of handling big data. Climate time series are often inputs to hydrologic models, which are widely used in water resources planning and management. To accelerate management of climate data, some researchers have developed standalone or web tools (Easterbrook and Johns, 2009; Martin et al., 2015; Wilby et al., 2002), which are often not very easy to use or easily accessible.

ii) Spatial downscaling and bias correction of GCM data are necessary before their use in regional impact analysis (Murphy, 1999; Chen et al., 2011). There are two distinct approaches to downscale GCM data; dynamic downscaling using a Regional Climate Model (RCM) and statistical downscaling. Dynamic downscaling is computationally expensive, and not always feasible to perform at the required spatial resolution, especially if predictions from multiple models are desired. Compared to the dynamical downscaling approach, statistical downscaling can be used to efficiently downscale a large number of GCM outputs to a fine temporal and spatial scale (Ahmed et al., 2013). The performance of statistical bias correction methods to downscale meteorological variables from GCMs is reported to be satisfactory in different hydro-climatological studies (Asong et al., 2016; Dettinger et al., 2004; Dosio and Paruolo, 2011; Hagemann et al., 2011; Ines and Hansen, 2006; Wilby et al., 2000).

GCMs often provide data at spatial resolutions larger than 50 km (~0.5°), which is too coarse to be directly used in local impact studies or regional planning (Hostetler et al., 2011). For a regional study that may contain only a few 0.5° grid points, spatial interpolation of meteorological variables could be used to produce a larger database. Many spatial interpolation techniques are developed (Di Piazza et al., 2011; Jeffrey et al., 2001; Ly et al., 2011; Wagner et al., 2012); Yang et al. (2015) compared four common interpolation techniques (ANUDEM, Spline, IDW, and Kriging) and concluded that inverse distance weighting method is slightly better than the others.

In this work, we combined the tasks of interpolation, downscaling, and bias correction together in one package, although there exist some studies and software that provide useful tools for some of these tasks separately. Santos et al. (2013) developed the Ultra-scale Visualization Climate Data Analysis Tools (UV-CDAT) to analyze and visualize climate data, especially the climate data of Earth System Grid Federation (ESGF). Rathjens et al. (2016) developed Climate Model data for hydrologic modeling (CMhyd). This tool extracts and bias corrects data from global and regional climate models and implements eight methods of bias correction such as linear scaling, delta change, distribution maps, etc. This program, however, uses an older version of CMIP3 climate data and has no capability to interpolate data to a finer resolution or to perform post processing of climate data to do extreme case analysis.

iii) Extreme hydrological events are often defined using diverse variables with different ranges in different climate zones. For example, drought may be defined in terms of meteorological, hydrological, agricultural, and socio-economic conditions. This has resulted in a large number of drought indices reported in the literature (Lloyd-Hughes and Saunders, 2002). Many studies have investigated occurrences of hydrological and meteorological extreme events using indices that focus on specific meteorological or hydrological events (Alexander et al., 2006; Bartolomeu et al., 2016; Beniston et al., 2007; Buckley and Huey, 2016; Goswami et al., 2006; Mishra and Singh, 2011; Mueller and Seneviratne, 2012; Orlowsky and Seneviratne, 2012; Panda et al., 2016; Rajczak et al., 2013; Schar et al., 2016; Um et al., 2016; Vicente-Serrano et al., 2012; Zhang et al., 2011).

As extreme thresholds vary for different climate and terrestrial zones, in CCT we provide a tool that can account for the occurrences of simultaneous extremes that have generated floods and droughts in the past in a given region. In previous studies, we used the concept of critical consecutive days (CCD) operator to calculate