Reconstruction of multiple climate variables at high spatiotemporal resolution based on Big Earth data platform

Mingxi Zhang

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Life Sciences, Faculty of Science

University of Technology Sydney

Australia

November 2021

Certificate of Original Authorship

I, Mingxi Zhang declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Life Sciences/Faculty of Sciences at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

Signature of student: Mingxi Zhang

Date: November 2021

Acknowledgements

This four-year Ph.D. journey has been an amazing experience for me that has changed my destiny. I would like to thank many people and institutions involved in the completion of this PhD study.

First and foremost, I would like to express my deepest gratitude to my research supervisors at UTS, Professor Qiang Yu, Dr. Xihua Yang and Professor Alfredo Huete, who have provided me generous guidance and support to accomplish this challenging project. Professor Qiang Yu, my principal supervisor, who is an inspiration in my life. He not only gave me an intellectually rich environment, but also gave me insightfulness guidance when having philosophical conversations with him. Without the help and support of Prof Qiang Yu, it would not be possible to get started my PhD study and bring this thesis to its completion. I am deeply grateful for the professional guidance of the principal research scientist Dr. Xihua Yang, who provided me immeasurable support and scientific motivation for my industry internship in NSW DPIE. Dr. Xihua Yang helped me with careful read and provided wealthy inputs to improve the theoretical contents of this thesis. I also greatly appreciate the help of Professor Alfredo Huete, who welcomed me to the group and expand my horizons on the state-of-the-art of technologies.

My special thanks go to the principal research scientist at NSW DPI, Dr. De Li Liu, who instilled in me the passion for scientific research and provided me with a great learning environment and invaluable support. I also owe my great gratitude to Dr. James Cleverly at UTS. I am so much in debt to his unconditional help. He helped me conceptualize the paper, instructed me on how to conduct scientific research and enrich it. Without his responsible step-by-step guidance, it would have been very difficult for me to make quick progress in my thesis.

I thank my colleagues from UTS, Dr. Bin Wang, Dr. Jie He, Dr. Puyu Feng, Dr. Qiaoyun Xie, Dr. Wenjie Zhang, and Dr. Cicong Gao, Dr. Song Leng, Dr. Lijie Shi, Dr. Rong Gan, Dr. Qinggaozi Zhu, Dr. Jianxiu Shen, Dr. Yuxia Liu, for your generosity and support, and for all the wonderful memories we shared. I would like to thank Professor Jiandong Liu, Dr. Jonathan

Grey, Dr. John Leys, Dr. Hongtao Xing, and Dr. Dongdong Kong for their insightful comments and support on my PhD study. I am also grateful to staff at UTS and NSW DPIE for their ongoing assistance in ensuring an enjoyable study environment and easy workflow during this wonderful PhD journey.

I would like to express my deepest gratitude to my beloved wife Dr. Hong Zhang and my parents in China. I thank for their perennial supported during my academic journey, and their encouragement has been a source of motivation for me to keep studying.

Finally, I would like to thank the China Scholarship Council and UTS for providing scholarships, and UTS and NSW DPIE for providing nice office environment to conduct this project.

Publications arising from this thesis

Journal papers directly related in this thesis:

Zhang, M., Wang, B., Liu, D. L., Liu, J., Zhang, H., Feng, P., ... & Yu, Q. (2020). Incorporating dynamic factors for improving a GIS-based solar radiation model. Transactions in GIS, 24(2), 423-441. (Chapter 2)

Zhang, M., Wang, B., Cleverly, J., Liu, D. L., Feng, P., Zhang, H., ... & Yu, Q. (2020). Creating new near-surface air temperature datasets to understand elevation-dependent warming in the Tibetan Plateau. Remote Sensing, 12(11), 1722. (Chapter 3)

Zhang, M., Yang, X., Cleverly, J., Huete, A., Zhang, H., & Yu, Q. (2021). Heat wave tracker: A multi-method, multi-source heat wave measurement toolkit based on Google Earth Engine. Environmental Modelling & Software, 105255. (Chapter 4)

Zhang, M., Yang, X., Leys, J., Gray, J., Zhu, G., & Yu, Q. (2020). The first combined water and wind erosion assessment for Australia 2000-2020. (Ready for submission) (Chapter 5)

Yang, X., Zhang, M., Oliveira, L., Ollivier, Q. R., Faulkner, S., & Roff, A. (2020). Rapid Assessment of Hillslope Erosion Risk after the 2019–2020 Wildfires and Storm Events in Sydney Drinking Water Catchment. Remote Sensing, 12(22), 3805. (Chapter 5)

Yang, X., Zhang, X., Lv, D., Yin, S., Zhang, M., Zhu, Q., ... & Liu, B. (2020). Remote sensing estimation of the soil erosion cover-management factor for China's Loess Plateau. Land Degradation & Development, 31(15), 1942-1955. (Chapter 5)

Research Grant directly related in this thesis:

Community vulnerability to extreme heat wave in Lake Macquarie area, Lake Macquarie Environmental Research Grants, 2019

Contents

List of Figures

List of Tables

Glossary

EO	Earth Observation
PB	Petabytes
GEE	Google Earth Engine
ODC	Open Data Cube
SH	Sentinel Hub
API	application programming interface
DEM	Digital Elevation Model
HPC	High-Performance Computing
CNN	convolutional neural network
DL	Deep Learning
STMSR	Spatial and Temporal Mountain Solar Radiation Modelling
SSR	Surface Solar Radiation
GSR	Global Solar Radiation
TP	Tibetan Plateau
LST	Land Surface Temperature
GWTR	geographically and temporally weighted regression
RF	Random Forest
XGBoost	eXtreme Gradient Boosting
LLTO	Leave-Location-Time-Out
LLO	Leave-Location-Out
LTO	Leave-Time-Out
FFS	Forward Feature Selection
EHF	Excess heat factor
HWN	Heat Wave Number
HWD	Heat Wave Duration
HWF	Heat Wave Amplitude
HWM	Heat Wave Magnitude
HWT	Heat Wave Tracker
GEV	Generalized Extreme Value
NEVA	Non-stationary Extreme Value Analysis
CMIP5	Coupled Model Intercomparison Project 5
WEPP	Water Erosion Prediction Project
SWEEP	Single-Event Wind Erosion Evaluation Program
RUSLE	Revised Universal Soil Loss Equation
RWEQ	Revised Wind Erosion Equation
GPM	Global Precipitation Measurement
GLDAS	Global Land Data Assimilation System
ERA5	European Centre for Medium-Range Weather Forecasts Reanalysis 5
SLGA	Soil and Landscape Grid of Australia
FVC	Fractional Vegetation Cover

Abstract

Reconstruction of climate variables with high spatio-temporal resolution is important when the meteorological observations required for environmental monitoring and modelling do not cover the study area. In addition, climate model reanalysis datasets suffer from coarse spatio-temporal resolutions, which fails to capture the complex variability of climate at fine scales. This thesis mainly reconstructed four climate datasets including: mountainous solar radiation, near-surface air temperature datasets over rugged terrain, five distinct metrics of long-term heat wave datasets, an updated database of water and wind erosion. For further use in practice, these datasets are freely accessible and online web application has been developed for academic research on climate change under accelerated global warming. The main findings of this thesis are:

(1) Incorporating dynamic factors for improving a GIS-based solar radiation model. Solar radiation has been a major input to agricultural, hydrological, and ecological modeling. However, solar radiation is usually influenced by three groups of dynamic factors: sun–earth position, terrain, and atmospheric effects. Therefore, an integrated approach to accurately consider the impacts of those dynamic factors on solar radiation is essential to estimate solar radiation over rugged terrain. In this study, a spatio-temporal gap-filling algorithm was proposed to obtain a seamless daily MODIS albedo dataset. A 1 km-resolution digital elevation model was used to model the impact of local topography and shading by surrounding terrain on solar radiation. A sunshine-based model was adopted to simulate radiation under the influence of clouds. A GIS-based solar radiation model that incorporates albedo, shading by surrounding terrain, and variations in cloudiness was used to address the spatial variability of these factors in mountainous terrain. Compared with other independent solar radiation products, our model generated a more reliable solar radiation product over rugged terrain, with an R^2 of 0.88 and an RMSE of 2.55 MJ m^{-2} day^{-1}. The improved solar radiation products and open-source app can be used further in practice or scientific research.

(2) Creating New Near-Surface Air Temperature Datasets to Understand Elevation-Dependent Warming in the Tibetan Plateau. The Tibetan Plateau has been undergoing accelerated warming over recent decades, and is considered an indicator for broader global warming phenomena. However, our understanding of warming rates with elevation in complex mountain regions is incomplete. The most serious concern is the lack of high-quality near-

surface air temperature (Tair) datasets in these areas. To address this knowledge gap, we developed an automated mapping framework for the estimation of seamless daily minimum and maximum Land Surface Temperatures (LSTs) for the Tibetan Plateau from the existing MODIS LST products for a long period of time (i.e., 2002–present). Specific machine learning methods were developed and linked with target-oriented validation and then applied to convert LST to Tair. Spatial variables in retrieving Tair, such as solar radiation and vegetation indices, were used in estimation of Tair, whereas MODIS LST products were mainly focused on temporal variation in surface air temperature. We validated our process using independent Tair products, revealing more reliable estimates on Tair; the R^2 and RMSE at monthly scales generally fell in the range of 0.9–0.95 and 1–2 °C. Using these continuous and consistent Tair datasets, we found temperature increases in the elevation range between 2000–3000 m and 4000–5000 m, whereas the elevation interval at 6000–7000 m exhibits a cooling trend. The developed datasets, findings and methodology contribute to global studies on accelerated warming.

(3) Heat wave tracker: a multi-method, multi-source heat wave measurement toolkit based on Google Earth Engine. Under ongoing global warming due to climate change, heat waves in Australia are expected to become more frequent and severe. Extreme heat waves have devastating impacts on both terrestrial and marine ecosystems. A multi-characteristic heat wave framework is used to estimate historical and future projected heat waves across Australia. A Google Earth Engine-based toolkit named heat wave tracker (HWT) is developed, which can be used for dynamic visualization, extraction, and processing of complex heat wave events. The toolkit exploits the public long-term high-resolution climate datasets to developed nine heat wave datasets across Australia for extreme heat wave value analysis. To examine climate change on heat waves and how they vary in time and space, we also explore the probability and return periods of extreme heat waves over a period of 100 years. The datasets, toolkit and findings we developed contribute to global studies on heat waves under accelerated global warming.

(4) The first combined water and wind erosion assessment for Australia 2000-2020 Soil erosion caused by water and wind is a complicated natural process that has been accelerated by human activity. This erosion has resulted in increasing areas of land degradation which threaten the productive potential of landscapes. Consistent and continuous erosion monitoring will help identify the trends, magnitude, and location of soil erosion. This information can then be used to evaluate the impact of land management practices and inform programs that aim to improve soil condition. We apply the Revised Universal Soil Loss Equation (RUSLE), Revised

Wind Erosion Equation (RWEQ), and an albedo-based wind erosion model to simulate water and wind erosion dynamics. With the advent of new or improved Earth observation big data, monthly and annual water, and wind erosion estimation at high spatial resolution (up to 90 m, 500 m) are produced for Australia from 2000 to 2020. We also evaluate the performance of three gridded precipitation products for rainfall erosivity estimation using ground-based rainfall. For model validation, water erosion products are compared with existing products and wind erosion results are verified with other models. We developed a water and wind erosion monitoring web application using Google Earth Engine. This web-based tool is particularly useful for identifying regions and specific locations where more sustainable land management practices should be encouraged.

Keywords: Big data; solar radiation; near-surface air temperature; heat wave; water and wind erosion; climate change; Cloud computing; China, Australia

Chapter 1. Introduction

1.1 Background to the question

Reconstruction of climate variables with high spatial and temporal resolution is important when the meteorological observations required for environmental monitoring and modelling do not cover the study area. In addition, climate model reanalysis datasets suffer from coarse spatial and temporal resolutions, which fails to capture the complex variability of climate at fine scales. Note that reconstruction of high spatio-temporal resolution climate data belongs to the domain of big data with large volumes of climate data producing from Earth observations and climate simulations. To address big climate data challenge, cloud computing and machine learning have been used for the building of climate data analysis tool and reconstruction of climate data, data modelling and prediction. As a background to support the research topic, this section will introduce big Earth Observation (EO) data, big EO meets climate change, big EO and Cloud Computing, big EO data with Machine Learning.

1.1.1 Big EO data

To understand big data, we must figure out what small data is. The term small data contrasts with big data, can be defined as small sets of data that are small enough to be conveniently stored, easily accessed and entirely processed on local servers or a laptop. On the other hand, big data refers to data too large and complex to be analysed and processed by traditional data-processing techniques in terms of volume, velocity, variety. In summary, big data also refers to extremely large chunks of structured, semi-structured and unstructured information, including: i) Structured data –transaction data, databases, and other structured data formats; ii) Semi-structured –system log files, text files etc.; and iii) Unstructured data – online data sources, digital images/audio/video feeds, sensor data, web pages, and so on (Pedamkar, 2020). If big data can be combined with small data approaches, the amount of information contained will be the greatest and the value embodied will be the highest.

Big EO data, as a subset of big data, is a hot topic and new engine for the climate system study. Big EO data involves extensive, accurate, continuous, and global data over the long-term period from the complex Earth system – atmosphere, biosphere, hydrosphere, lithosphere, and cryosphere. Such observation data can help reveal the explicit and implicit spatio-temporal processes occurring at the Earth's surface (Guo, 2017).

Big EO data in climate research differs from other types of big data in terms of characteristics and attributes (Guo, 2017). In general, big EO data in climate research has all the feature of big data but has at least the following four distinct features: (1) With respect to the data volume or data update rate, e.g., Petabytes (PB) of EO data have been acquired and stored globally at an accelerated rate. (2) As to data variety or data acquisition approaches, big EO data is multi-sourced, generally from in situ observations, satellite observations, and Earth system model simulations. (3) In respect to spatio-temporal characteristics, the explosively growing big EO data is multi-dimensional, differs in a variety of spectral, spatial, and temporal resolutions. (4) Regarding data-driven analysis methods, big EO data tends to use statistical methods and high spatio-temporal resolution Earth system models for scientific discovery (Zhang and Li, 2020).

Due to the four features of big EO data, climate system research is also facing the challenges, which are briefly reviewed as follows: (1) When it comes to data storage, it fails to storage big EO data on traditional hard disk drives and is problematic with planetary-scale storage. For example, DigitalGlobe currently archives 70 PB of satellite imagery. (2) When transferring big EO data from scientific data centres to local host is also challenging, e.g., there are 40+ years of remotely sensed data available from a wide range of satellites and sensors. (3) Managing big EO data (describing, cleaning, storing, and organizing) efficiently is even more challenging due to the data's spatio-temporal characteristics. (4) Analysing big EO data challenges the complexity and scalability of analysis/mining algorithms. (5) It is difficult to provide real-time and human-interactive visualization to analyse and explore big EO data.

Thus, there is a need for huge technological progress in big EO data, even disruptive

changes to address these challenges. For big EO data storage, hosting big EO data on cloud storage services (e.g., Google Cloud Storage, Amazon Elastic S3, Microsoft Azure), which can scale up quickly, provides solutions to address big EO data storage challenges. For data transmission and processing, the solution is to allow software to process big EO data online without downloading data or move computation to data. In addition, smart data compression algorithms and pre-processing techniques are recommended. For big EO data analysing, the gap can be filled by welding analysis programs to cloud computing platform and developing new tools to harness the distributed processing power. For real-time and human-interactive visualization, an envisioned workflow is to produce analysis-ready data in cloud-based platforms where EO analysts can easily access and explore them with their own or already existing tools. The end tools and data product will also reside within the cloud-based platforms, which can be shared and accessed by the other end users. The interactive components of the tools might be the combination with geographical or temporal filters, colour and data source selection (Sudmanns et al., 2019).

1.1.2 Big EO data meets Climate Change

Climate change as a data-intensive subject has been the research focus of big data scholars over the past several decades. With the help of big EO data, it is feasible to know what is causing the climate and environmental changes in our planet. We can even utilize large-scale and long-term time series EO data to better predict the future and provide more viable measures for dealing with potential climate hazards. Big data in climate change has two fundamental elements: the big EO data resources and the big EO analytics techniques. It is well-known that climate system models provide a new and dynamic way to assess past, present, and future climate and environmental change. However, the grid size of cells in current climate simulations is at various coarse levels ranging from 25 km to 2 degree. By contrast, high-resolution climate simulations (less than 5 km) will help regional climate adaptation, improve forecasts of extreme weather, and even assess climate warming in mountainous regions like Tibetan Plateau. By downscaling the size of grid cells of climate system models to 1 km or less,

critical and accurate components of the atmospheric and oceanic could be precisely modelled. However, the increased realism of high-resolution climate simulations comes at a cost. When the resolution increases, the simulation of climate models faces the challenge of having to collect, characterize, and analyse large amount of data, also needs to consider the multi-source, multi-variable, and multi-scale data with the different spatial and temporal attributes. For example, data outputs from climate forecasting models are updated hourly and the amounts have reached more than 300 TB per day. Therefore, it is necessary to bring computing and data together by no longer moving data to computing but computing to data. Therefore, it is foreseen that the data processing and analytical capabilities associated with the cloud and distributed computing paradigms are a crucial part of future climate modelling. However, processing and analysing these big EO data also face many difficulties. Improved Machine Learning algorithms based on big EO data provide another new way to help scientists find patterns and make predictions. This is especially true for big EO with a very large number of variables (Zhang and Li, 2020).

1.1.3 Big EO data and Cloud Computing

The tremendous increase in Big EO data has posed grand challenges for the data management such as data storage, post-processing, analytics, online visualization, sharing, and applications. However, the emergence of cloud computing provides critical computing support to meet these challenges (Li et al., 2020b). Cloud computing platforms are efficient ways to access, analyse and store big EO datasets on supercomputers, providing the users with infrastructure, platform, storage services, and software packages (Amani et al., 2020).

This study will give an overview of three widely used cloud computing platforms for big EO data in terms of data, features and available capabilities: Google Earth Engine (GEE), Open Data Cube (ODC), and Sentinel Hub (Gomes et al., 2020). The free-to-use GEE platform provides access to i) petabytes of publicly available EO datasets and other ready-to-use scientific products; ii) high-speed parallel computing capabilities, the state-of-the-art machine learning algorithms; and iii) a geoprocessing Application Programming Interfaces (APIs) library with a development environments that supports popular coding languages such as JavaScript and Python and extensive education tutorials (Tamiminia et al., 2020).

Sentinel Hub (SH) is also a big EO data management and analysis platform that provides cloud-based application programming interface (API) for global archive of EO data processing. However, SH is a private, fee-based cloud computing platform, but is available for limited public access (https://www.sentinel-hub.com). The SH platform contains: i) Archives of more than 5 PB of EO data including Landsat, MOIDS, ENVISAT, and Sentinels are accessed over the web application; ii) Powerful multi-temporal remote sensing change detection and land cover classification data analysis; iii) Evalscripts, representational state transfer (REST) interfaces and open-source libraries are provided for developers to build new application. By contrast, ODC is an open-source framework consisting of a set of spatio-temporal data structures and geographic analysis tools, which aims to index, manage, and analyse big EO data. ODC accesses and manipulates big EO datasets through a set of command line tools and Python API.

Overall, GEE is the platform that delivers the best solution for users in terms of ease of use and development maturity. However, it has limitations because it is a closed business platform, especially as to ensure the sustainability, scalability, and reproducibility of the usage. For more complex scientific analyses, platforms like ODC could be allowed to scale through open-source framework so that scientists has direct access to powerful processing capabilities of cloud computing infrastructure (Gomes et al., 2020).

1.1.4 Big EO data with Machine Learning

With the advent of the big data era, the multi-source, multi-dimensional and multi-scale meteorological data has become typical big EO data collection with spatio-temporal structure. Traditional approaches may not be optimal in multi-dimensional time series weather forecast, whereas Machine Learning (ML)/Deep Learning (DL) approaches are able to extract spatio-temporal context and gain better understanding of weather system behaviour. Applications of ML/DL to the climate, such as for weather forecast and climate change, has led to important developments. A lot of ML algorithms and their variants have been now extensively used in the climate change literature. Currently, RF is the most popular one for classification and regression purposes. Alternative ML algorithms such as artificial neural network, support vector machines, partial least squares regression are also widely used. The literature shows that ensemble modelling can get better performance and the higher estimated performance of ML algorithms is not only affected by the optimization of hyperparameters, but also partly depends on the validation strategy (Bonavita et al., 2021; Meyer et al., 2018).

In recent years, DL has produced outstanding results in forecasting many Earth system components including climate predictions. DL is appropriate to mine complex spatial and temporal relationships between meteorological data elements, and DL has been envisioned as a promising research topic to cope with the big EO data challenges faced by traditional theory-driven approach (Ren et al., 2021).

This study will survey the state-of-the-art DL methods for climate forecast. We summarize the basic DL models, Hybrid DL models and Coupling DL and physical models. The selection of basic DL models is based on big EO data characteristics. The typical Autoencoder-based DL models are suitable for noisy reduction of climate data with high-dimensional (Hossain et al., 2015). The convolutional neural network (CNN) models are extensively applied for image processing that extreme weather phenomena (e.g., typhoon, rainstorm, atmospheric rivers) can be detected (Ham et al., 2019). The long short-term memory (LSTM) models are used for climate prediction that contains long time sequence (Shi et al., 2015). The hybrid DL models compose of the basic deep neural network (DNN) models to capture spatial and temporal structures of meteorological datasets (Chen et al., 2019). Hybrid DL models can be grouped into two categories: one belongs to the spatio-temporal sequence weather prediction, the other is for classification and pattern recognition with extreme weather detection. A typical hybrid DL architecture contains three parts: the input, hidden and output layers. The input layers contain meteorological attributes as the input for DL models. The hidden layers consist of two components: the CNN part with convolution, pooling, flatten and fully connected layers is used for capturing spatial features and correlation, while LSTM part is used for capturing temporal features and correlations. The output layers are referred to desired climate forecast. Considering the respective merits of physical approaches and DL models, bridging two paradigms has recently been envisioned as an attractive research topic (Jiang et al., 2020). The existing approaches to coupling models can be summarized into four groups: 1) To train DL models with prior knowledge from physical models; 2) To constrain predictors of DL models with physical laws; 3) To replace the empirical subprocess parameterizations using DL models; 4) To incorporate ordinary and partial differential equations into DL models.

After the summarization of ML/DL methods, we will present the drawback of ML/DL in three aspects: computability, generalization, and interpretability. In terms of computability, most ML/DL models are typically trained with GPUs and TPUs, which heavily depends on high-performance computers even supercomputers. Second, the generalization of ML/DL

models is limited, i.e., it is difficult to apply the trained ML/DL models for large-scale and long-range climate forecast because many hyperparameters need to be tuned and optimised. Finally, interpretability has been identified as a generic drawback of complex ML/DL models. By contrast, a simple model enables interpretation and visualization of the model simulation and prediction explanations. Meanwhile, the higher accuracy of complex ML/DL models comes at the cost of providing meaningful explanations and causal links between covariates and predicted values.

1.2 Statement of significance and knowledge gaps

Solar radiation is the primary driving force for earth system processes, and its supply is a major input to agricultural, hydrological, and ecological models (Aguilar et al., 2010; Brock, 1981; Fu and Rich, 2002). Additionally, existing solar radiation products are mostly at coarse resolution (greater than 10 km grid spacing). Therefore, fine spatial and temporal mapping and monitoring of solar radiation components are essential for the design in solar energy systems. However, a GIS-based solar radiation model that allows for the treatment of high spatial and temporal variability in sun-earth position, terrain and atmospheric effects has not yet been developed for monitoring daily solar radiation. Much effort therefore needed to build a computationally economical, next generation GIS-based solar radiation model, which could explain influential impacts from albedo, surrounding terrain and cloud.

The Tibetan Plateau (TP) is named the "the third pole of the Earth", the highest and largest plateau globally (Qiu, 2008). The TP exerts profound dynamical and thermal influences on the regional and global climate (Duan et al., 2012; Manabe and Terpstra, 1974). For global warming, TP is considered as an early warning sign. Over the period of 1984-2009, TP has undergone serious warming, with a warming rate of 0.46 °C decade^{-1}, which is almost 1.5 times the rate of global warming (Kang et al., 2010; Kuang and Jiao, 2016). Accelerated warming on the TP has intensified permafrost degradation, snow melt and glacier retreat (Yang et al., 2014). Presently, the status of TP warming is evaluated through the analysis of Tair at meteorological stations. However, most meteorological stations are located in the eastern TP below 3800m. Because of sparse high-elevation meteorological observations in central and northwest of TP, there is a possibility that our understanding of warming rates with elevation in complex mountain regions is incomplete. In addition to limited coverage by in-situ measurements, Tair at TP suffers from extreme local variability due to factors such as topography and exposure

(Pepin et al., 2015). Moreover, the Himalaya mountains only reach heights of about 6,000 m in latest simulated 9 km grid climate products. Therefore, improved Tair estimations by developing high-resolution near-surface air temperature datasets considering rugged terrain over the TP is a crucial step for understanding the accelerated warming in the TP.

Under ongoing global warming due to climate change, heat waves are expected to become more frequent and severe in the future. Extreme heat waves during the last two decades have been recorded across many regions in the world such as those in Europe in 2003 (Schär et al., 2004), Moscow region in Russia in 2010 (Rahmstorf and Coumou, 2011), and Australia in 2013 (Lewis and Karoly, 2013). Although a heat wave is commonly known as a period of exceptional hot weather event, there is currently no universal informative measurement in climate science community (Alexander and Perkins, 2013). To overcome these issues, a set of climate indices developed by the Expert Team on Climate Change Detection and Indices (ETCCDI) have been widely applied to observational and modelled climate data to understand previous and future changes in extreme heat wave events (Alexander et al., 2006; Zhang and Yang, 2004). The work by ETCCDI is extensively recognized as pioneering, however, the indices only measure one feature of extreme events such as frequency, intensity or duration (Perkins, 2015). A comprehensive and consistent analysis of heat waves is required, which should consider multi-characteristics of heat wave events, namely: i) frequency, ii) intensity, iii) duration, and iv) spatial extent (Raei et al., 2018).There is no general heatwave measurement package which has an imperative advantage of applying big climate data at fine spatiotemporal scale. Desktop MATLAB toolbox like Global Heatwave and Warm-spell Data Record and Analysis Toolbox (GHWR) still has a bottleneck when encountering the challenges related to accessibility of long-term gridded climate data, data storage and computational requirements. In the current era of big spatial and Earth Observation (EO) data, users need to deal with a vast amount of different spectral, temporal and spatial resolutions data (Gomes et al., 2020). To meet these demands, there is need for novel technologies based on cloud computing to properly extract heat wave information at the server side without having to download vast amounts of climate data and provide dynamic visualization, extraction, and processing of complex heat wave events.

Soil erosion is a major threat to sustainability of agriculture (Borrelli et al., 2017; FAO, 2015). Under changing land use and climate, soil erosion from water and wind has accelerated with resulting economic, social, and environmental implications, both on-site and off-site (Telles et al., 2013). On-site, water and wind erosion causes the loss of soil, nutrients and

organic matter that results in decreased soil fertility and land productivity (Zhang et al., 2019). The reduced productivity of farmland means that about 10 million ha of cropland worldwide is abandoned yearly due to soil erosion(Chappell et al., 2019; Faeth, 1994). This further leads to reduce the social viability and population levels in rural communities, influencing long-term sustainable regional development. The subsequent sedimentation and nutrient loss may also cause off-site environmental, air (Middleton, 2019) and water quality degradations. In Australia, for example, the assessment Bui et al. (2010) concluded that soil erosion in Australian cropping regions was occurring at unsustainable rates and has a critical impact on agricultural productivity. Environmental impacts of excessive sedimentation and nutrient delivery on inland waters, estuaries and coasts are already occurring. The net median erosion rate in cultivated regions is estimated 1.26 Mg ha^{-1} yr^{-1} (Chappell et al., 2011), and 7% of Australia had soil losses of more than 1 Mg ha^{-1} yr^{-1}. It also should be noted that Australia is the most fire-prone regions of the world. Wildfire related water erosion in Australia is responsible for reef deterioration, roads damage, river pollutants (Yang et al., 2020). In addition, wind erosion from arid and semi-arid areas of Australia severely affects the air quality in the coastal zone where most Australians live (Leys et al., 2011). Since 2000, the millennium drought and mega-fires in Australia also prompt the urgent need to revisit soil erosion dynamics to provide a more contemporary view of water and wind erosion trends.

1.3 Research issues and objectives

Scientific data is one of the prerequisites for conducting climate research. The key to the scientific questions in climate research is to utilize various sensors to obtain accurate and critical climate variables. As the demand of high-quality climate and environmental datasets to support agricultural climate services, weather hazard risk reduction, and climate change adaptation and mitigation continues to grow, it becomes particularly important to combine observed climate information with climate simulation outputs that can produce more accurate weather datasets at a given time. Observational climate information includes both ground stations and big EO data. For the ground station data, various geostatistical interpolation approaches, such as kriging interpolation and spline function methods, are commonly applied to obtain spatially continuous climate data. In contrast, the processing of big EO data is an

upward spiral process. Only by continuously reprocessing big EO data with the latest technology can the quality of satellite climate data sets be continuously improved until they fully meet the needs of climate change research.

Big data and climate research are closely related, many climate research studies cannot be done without big data. Climate change models require computational resources with large amounts of storage and fast access to ever-increasing amounts of data. With the massive amount of climate data being generated, user-friendly cloud-based software and platforms are needed to visually manage and display the data. Furthermore, this study will use mechanistic models and statistic methods to provide high-quality long-term climate datasets for China and Australia. This project will try to provide answers to the following questions:

(1) What is the spatio-temporal solar radiation variation over the Loess Plateau from 2003 to 2014?

(2) What is the rate of warming above 5000 m elevation at Tibetan Plateau?

(3) How does the extreme heat wave risk change in Australia recent climates and future climate conditions?

(4) What are the trends in soil erosion by water and wind across Australia since 2000?

The specific aims of this study are to:

(1) To develop an improved GIS-based solar radiation model that allows treatment of high spatial and temporal variation of albedo, surrounding terrain shading and cloud to monitor daily solar radiation at fine resolution.

(2) To develop an automated mapping framework for the estimation of combined and seamless Terra and Aqua MODIS Land Surface Temperatures (LST) for the global. The machine learning models combined with MODIS LST and meteorological station data to provide reliable temperature products at high-resolution over the rugged mountainous area.

(3) To develop a multi-method global heat wave data record and analysis toolbox (namely Heat Wave Tracker) to process and extract heat wave records from multi-source climate datasets.

(4) To develop a water and wind erosion monitoring web application toolbox to estimate monthly and annual soil loss by water and wind across Australia from 2000 to 2020.

1.4 Thesis outline

The proposed outline for this thesis is as follows:

(1) Introduction

(2) Incorporating dynamic factors for improving a GIS-based solar radiation model

(3) Creating new near-surface air temperature datasets to understand elevation-dependent warming in the Tibetan Plateau

(4) Heat wave tracker: a multi-method, multi-source heat wave measurement toolkit based on Google Earth Engine

(5) Assessment of soil erosion by water and wind for Australia 2000-2020

(6) Final conclusions and future research

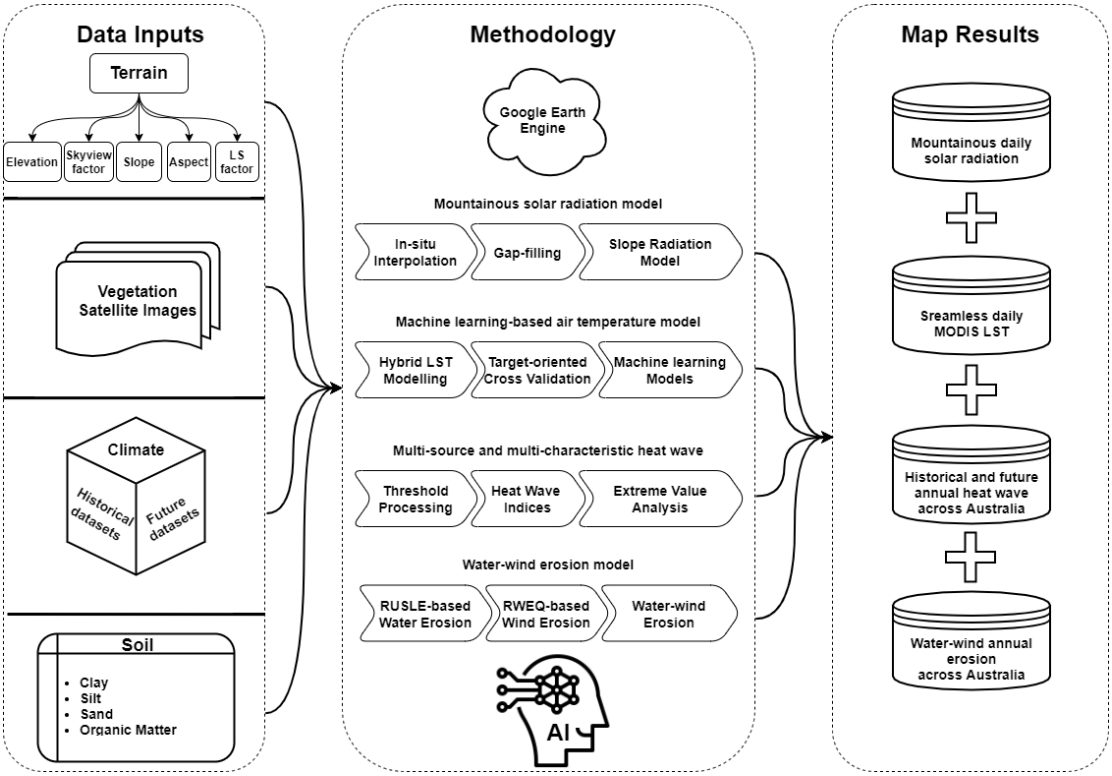


Figure 1-1 The framework of the thesis.

Chapter 2. Incorporating dynamic factors for improving a GIS-based solar radiation model

This chapter is based on the following manuscript:

Zhang, M., Wang, B., Liu, D. L., Liu, J., Zhang, H., Feng, P., ... & Yu, Q. (2020). Incorporating dynamic factors for improving a GIS-based solar radiation model. Transactions in GIS, 24(2), 423-441.

Abstract

Solar radiation has been a major input to agricultural, hydrological and ecological modeling. However, solar radiation is usually influenced by three groups of dynamic factors: sun-earth position, terrain, and atmospheric effects. Therefore, an integrated approach to accurately consider the impacts of those dynamic factors on solar radiation is essential to estimate solar radiation over rugged terrains. In this study, a spatial and temporal gap-filling algorithm was proposed to obtain seamless daily MODIS albedo dataset. A 1km-resolution DEM was used to model the impact of local topography and of shading by surrounding terrain on solar radiation. A sunshine-based model was adopted to simulate radiation under the influence of clouds. A GIS-based solar radiation model that incorporates albedo, shading by surrounding terrain and variations in cloudiness was used to address the spatial variability of these factors in mountainous terrain. Compared with other independent solar radiation products, our model generated a more reliable solar radiation product over rugged terrain, with an R^2 of 0.88 and an RMSE of 2.55 MJ m^{-2} d^{-1}. The improved solar radiation products and open-source app can further be used in practice or scientific research.

Key words: *Solar radiation modeling, DEM, MODIS albedo, gap-filling algorithm, rugged terrains, Opensource, GIS-based solar radiation model*

2.1 Introduction

Solar radiation is the primary driving force for earth system processes, and its supply is a major input to agricultural, hydrological, and ecological models (Aguilar et al., 2010; Brock, 1981; Fu and Rich, 2002). Therefore, knowledge of the spatial and temporal variability of incoming solar radiation is critical for understanding these processes. Additionally, fine spatial and temporal mapping and monitoring of solar radiation components are essential for the design in solar energy systems.

The spatial and temporal heterogeneity of solar radiation over rugged terrain is determined by three groups of dynamic factors: sun-earth position, terrain, atmospheric effects (Pintor et al., 2015). Based on the sun-earth geometry formulation, the first group can be precisely calculated. For the other two groups, the effects of terrain (shadowing, absorption, and reflection) and atmosphere are difficult to model due to their dynamic nature. Particularly, albedo of the underlying surface modulates the amount of solar radiation absorbed and reflected by that surface and directly controls the distribution of solar radiation between the surface and the atmosphere. Additionally, shadows cast by complex topography due to different incident angles of the rays determine the fraction of direct and diffuse radiation in global solar radiation. Furthermore, clouds play a major role in the atmospheric attenuation of incoming solar radiation, but modeling of the radiative effects of clouds is challenging due to their variability in time, location and condition. Hence, quantitative modeling of the impacts of those dynamic factors on solar radiation is essential to accurately estimate solar radiation over rugged terrain.

Three major methods have been used for solar radiation modeling over the past few decades, namely traditional interpolation methods, GIS-based solar radiation models and satellite-derived solar radiation estimates (Hofierka, 2002; Qin et al., 2015; Ruiz-Arias et al., 2009; Zhang et al., 2015). In spatial interpolation methods, unknown values are reliably predicted from ground-based measurements and external complementary information. However, the reliability of such methods strongly depends on sample size and the complexity of the topography (Alsamamra et al., 2009). By contrast, GIS-based solar radiation models

(Table 2-1) such as Solar Analyst (Fu and Rich, 2002), SRAD (Wilson, 2000), Solei-32 (Mészároš and Miklánek, 2006) and r.sun (Hofierka, 2002) have been developed to calculate the incoming solar radiation for each cell of a DEM (Digital Elevation Model) during recent decades. These GIS-based models are technologically interoperable and scientifically rigorous, but they use different algorithms (either physically-based or empirically-based), thus their results show large differences in estimating solar radiation (Ruiz-Arias et al., 2009). Two limitations of these GIS-based solar radiation models are that they are computationally demanding and that they have difficulty incorporating dynamic factors that contribute to solar radiation estimates (Freitas et al., 2015). In particular, Solar Analyst is a GIS-based sun-earth geometric model, but it ignores reflected radiation from nearby surfaces. However, accounting for reflected radiation is vital at locations with high albedo due to snow-cover because any variation in snow-cover albedo can have a great impact on solar radiation (He et al., 2014). Unlike Solar Analyst, SRAD estimates reflected radiation, but its reliability declines when monthly average cloudiness and sunshine hours are used to adjust daily shortwave radiation. Furthermore, processing of large-scale DEMs is not appropriate using Solei-32, Solar Analyst or SRAD, all of which suffer from heavy computation demand with very large datasets (Tabik et al., 2012). Additionally, both Solei-32 and r.sun require appropriate parameters for estimating the atmospheric attenuation of incoming solar radiation, such as atmospheric transmissivity, the circumsolar coefficient, and atmospheric turbidity. However, vertical profiles of many atmospheric parameters are rarely available, especially in mountainous areas. Even when atmospheric parameters are available for these GIS-based tools, they consider only shelter effects due to the slope, whereas effects of the surrounding topography should be taken into account (Wang et al., 2014a). Recent studies have found that satellite-based solar radiation estimates provide reasonable values and large spatial coverage. One weakness of satellite-based estimates resides in cloud detection, where even a small cloud can make solar radiation estimates less accurate. In addition, the accuracy of satellite-based solar radiation estimates for complex topography is still limited (Romano et al., 2018; Roupioz et al., 2016; Tang et al., 2016; Yeom et al., 2016).

Much effort therefore needed to build a computationally economical, next generation GIS-based solar radiation model, which could explain influential impacts from albedo, surrounding terrain and cloud. However, a GIS-based solar radiation model that allows for the treatment of high spatial and temporal variability in sun-earth position, terrain and atmospheric effects has not yet been developed for monitoring daily solar radiation. In recent years, the advanced cloud-based geospatial computing platform, Google Earth Engine (Gorelick et al., 2017b), has given researchers the opportunity to use big data for planetary-scale environmental data analysis. The present study covers this gap by complementing existing solar radiation studies with a dynamic spatial perspective, by incorporating the spatial heterogeneity of factors into a model and by applying cloud-based geospatial computing techniques to the problem.

In this study, DEM and land surface albedo data were used to determine whether each point in the landscape was shaded by surrounding terrain. A generic spatial and temporal gap-filling algorithm was then developed to retrieve seamless albedo datasets from the raw MODIS product (see Methods). It is worth noting that other remote sensing indices with missing values can also be gap-filled using this algorithm. A sunshine-based submodel was used in this study as a module to address actual radiation under the influence of clouds. An assessment of overall model accuracy was made by comparing our modeling results with ground observed data and existing solar radiation products. The GIS-based model developed in this study has been released to the research community in a publicly available online platform, the spatial and temporal mountainous solar radiation model (STMSR), after comparison with current GIS-based solar radiation modeling software. This online mountainous solar radiation model can be extended to other locations with around the world complex terrain.

Table 2-1 A comparison between the proposed tool and related tools

DEM-based Model	Environment	Computing capacity	Ground parameter	Atmospheric parameter	Sky view factor
STMSR	GEE	Cloud-based unlimited	Dynamic albedo	Dynamic cloud	Surrounding terrain
r.SUN	GRASS	Multi-	Dynamic	Static	Slope itself

		processor limited	albedo	coefficients	
SRAD	ArcGIS	Multi-processor limited	Static albedo	Static coefficients	Slope itself
Solar analyst	ArcGIS	Single-processor limited	Not included	Static coefficients	Slope itself
Solei-32	DOS	Single-processor limited	Static albedo	Static coefficients	Slope itself

2.2 Materials and methods

2.2.1 Study area and observed solar radiation data

The Loess Plateau is a 64 million hectare, semi-arid region located in north-central China (33° 43' to 41° 16' N and 100° 54' to 114° 33' E) (Lü et al., 2012). The Loess Plateau has irregular topography with varying elevation between 422 m and 3390 m above mean sea level (Figure 2-1). Studying the topography impact on solar radiation is of major importance on the Loess Plateau because of its distinct variation in topography. The Loess Plateau's extensive landscape is diverse. At the local scale, the terrain in the Loess Plateau includes eroded gully, near-vertical slopes, varying terraces, shoulders, and summits. For macro landforms, the diverse topography contains high mountains, rough hills, broken tablelands, and low plains. This region has played an increasingly important role in China's ecological security and natural resources supply (Zhao et al., 2013). Since the ecological restoration projects such as "Natural Forest Protection" were implemented in this area, sloping cropland was converted to orchard land, and forest land has increased significantly. Simultaneously, there has been accelerated warming in the southwest region of the Loess Plateau (Sun et al., 2015).

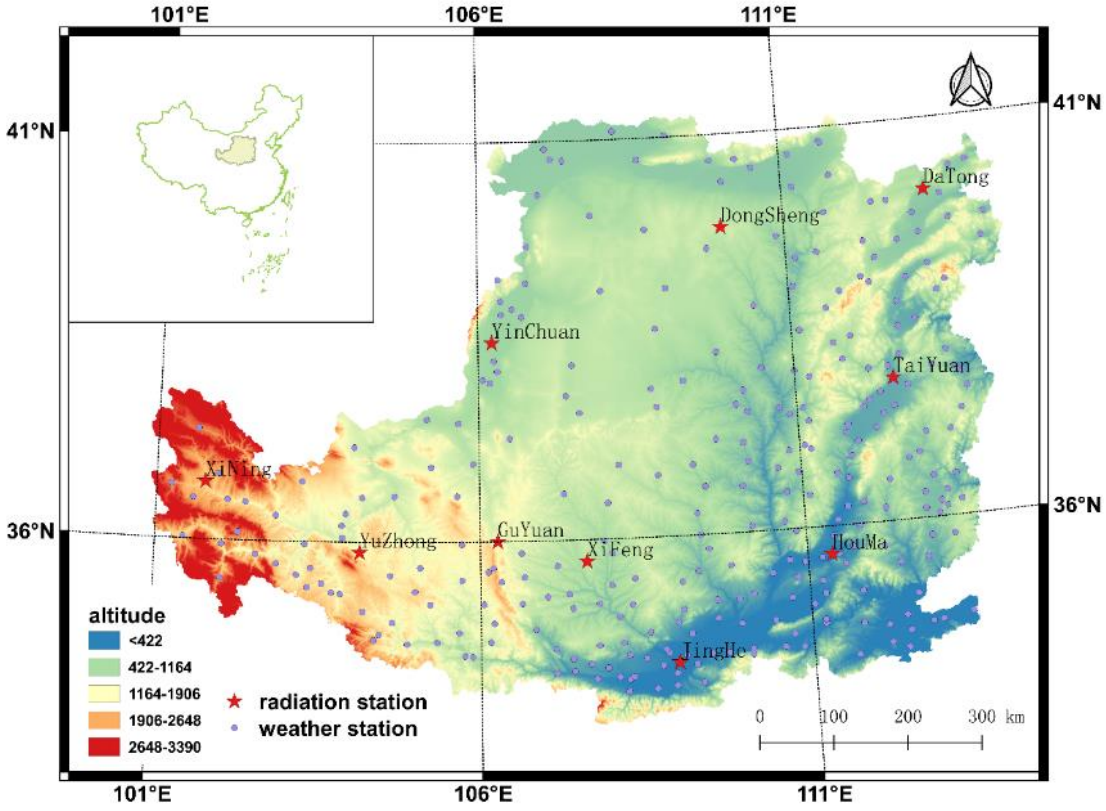


Figure 2-1 The study area showing the Loess Plateau located in north-central China including 10 radiation stations and 301 weather stations.

We acquired data from 301 Loess Plateau weather stations, carefully examined the data for quality and removed null values, and ingested the data into cloud storage. Other relevant data sources were DEM data and MODIS surface albedo from the cloud data catalog, i.e., SRTM Digital Elevation Data 90m (Farr et al., 2007). During ingestion, DEM data were stored at various levels of resolution, from native resolution (90 m) to increasingly coarse levels. This was done by aggregating data in a pyramid structure such that pixel values of an upper level are the mean of pixels at the next lower level. The resolution of the DEM used for calculation was the closest scale equal to or less than the scale of the data source with the coarsest native resolution in our analysis. The coarsest resolution in our study was obtained from the daily MCD43A3.006 product, which provides 1 km resolution for black-sky and white-sky albedo across each of the MODIS surface reflectance bands (from band 1 to band 7) as well as three broad-spectrum bands (Schaaf, 2015).

GLDAS assimilates satellite and ground-based observational data products (Rodell et al., 2004a) to generate land surface parameters. This dataset supports agricultural and meteorological modeling. The GLDAS dataset started on January 1, 1948 and continues to the present time. The temporal and spatial resolution is 3 hours and 0.25 degrees, respectively.

Land process research requires high spatial and temporal forcing data of Surface Solar Radiation (SSR), which was derived by the fusion method of MODIS and MTSAT (Tang et al., 2016). MTSAT data includes MTSAT-1R and MTSAT-2, obtained from the Japanese Meteorological Agency. The temporal resolution of a MTSAT image is 30 min. A MTSAT image has five channels, and the spatial resolution for the visible sensor at nadir is 1 km, and for the other infrared sensors is 4 km. SSR was estimated by combining signals of polar-orbit (MODIS) and geostationary satellites (MTSAT).

Table 2-2 Data sources for calculating and comparing solar radiation

Sequence number	Data name	Data time span	Data source
1	Sunshine data	2005-2014	China Meteorological Administration
2	DEM	2010	USGS/SRTMGL1_003
3	MODIS Albedo	2000-2017	NASA LP DAAC at the USGS EROS Center
4	Surface Solar Radiation	2007-2014	Third Pole Environment Database
5	GLDAS2.1	1979-2018	NASA

2.2.2 Schematic of the modelling

The GIS-based solar radiation model developed in this study (STMSR, the spatial and temporal mountainous solar radiation model) can also be seen as a DEM-based model that integrates with a geospatial cloud-based computation platform to simulate the dynamics of solar radiation. The required inputs for the model include a DEM, MODIS albedo, in situ observational data and empirical coefficients. The entire process of modeling solar radiation in a mountainous terrain includes the three steps shown in Figure 2-2. The first step was to

estimate extraterrestrial solar radiation and sky view factor on slopes in a High-Performance Computing (HPC) environment by using parallel raster image processing before uploading those image datasets into the cloud data catalog. The second step was to retrieve horizontal solar radiation data, including global solar radiation, direct solar radiation and diffuse solar radiation and gap-filled MODIS albedo. The third step was to build a spatial and temporal mountain solar radiation model with those input parameters and create an online spatial and temporal mountain solar radiation modeling app.

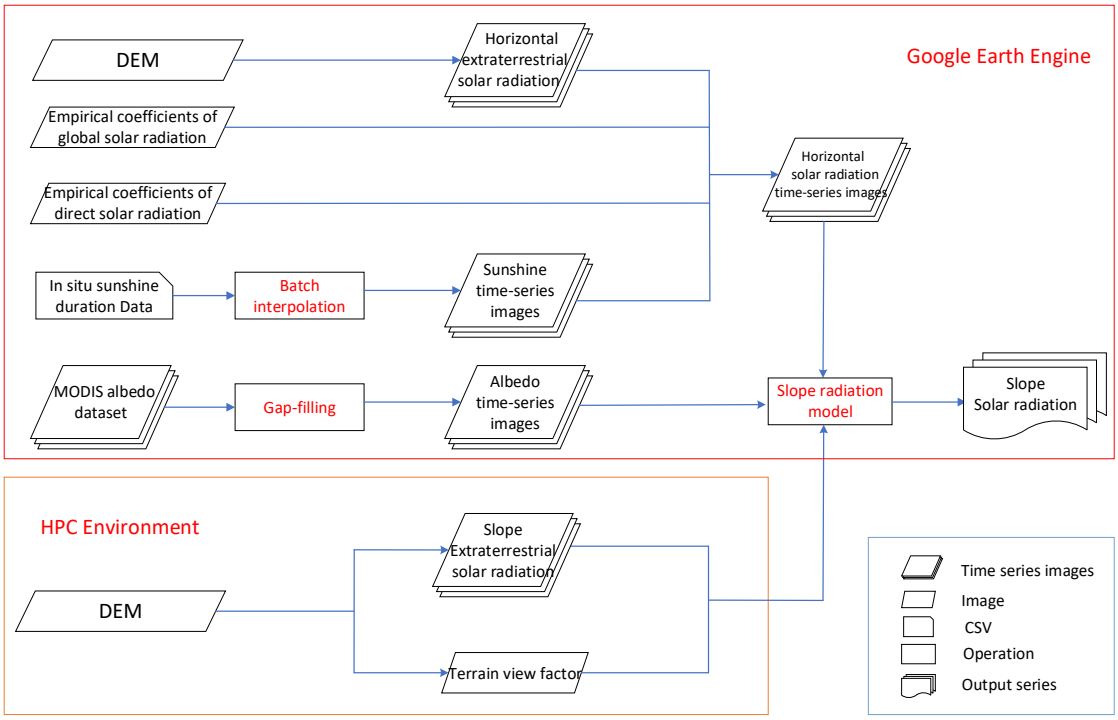


Figure 2-2 Flowchart of steps for calculation of solar radiation in mountainous terrain.

2.2.3 Distributed Global Solar Radiation (GSR) model for rugged terrain

2.2.3.1 Radiation on the horizontal surface

As reflected radiation on a horizontal surface is negligible, the radiation on a horizontal surface is partitioned into two parts, the beam and diffuse radiation, which are usually estimated by statistical regression of observed data(Liu et al., 2009).

$$K_b = B_h/Q_h \ , \tag{1}$$

$$K_d = D_h/Q_h \ , \tag{2}$$

$$K_b + K_d = 1, \tag{3}$$

where the direct radiation fraction K_b is called the direct radiation transmittance, and K_d is named the diffuse radiation fraction.

Since clouds are dynamic and site-specific, much observational data is required to parameterize cloud effects. Observations of routine meteorological variables such as sunshine and temperature do not require complicated instruments. A sunshine-based submodel is used in this study, because it produces better solar radiation estimates than cloud-based or temperature-based models (Iziomon, 2001; Podestá et al., 2004; Trnka et al., 2005). For example, the major limitation of cloud-based models is that they show systematically larger differences between measured and modeled values as cloud cover increases (Trnka et al., 2005). K_d is derived as a polynomial function of sunshine duration (Iqbal, 1983), Q_h is often derived from sunshine duration percentage using the Ångström formula (Angstrom, 1927). B_h is a polynomial function of relative sunshine duration (Louche, 1991). A further step is that Zeng et al. (2005) established an exponential function of direct radiation and global horizontal radiation.

$$Q_h = (a_h + b_h \times s) \times Q_{sh}, \tag{4}$$

$$B_h = (1 - a) \times (1 - e^{\frac{-bs^c}{(1-s)}}) \times Q_h, \tag{5}$$

where Q_h is the horizontal solar radiation (MJ·m^{-2}·d^{-1}), Q_{sh} is the horizontal extraterrestrial radiation (MJ·m^{-2}·d^{-1}), s is the relative sunshine duration (i.e., the ratio of daily bright sunshine duration to the maximum possible duration of sunshine in daylight hours), a_h, b_h, a, b, and c are regression coefficients. The coefficients a_h and b_h in the Ångström equation were calibrated individually for each station in China using monthly observations. The calibration of direct radiation coefficients was achieved using least square linear regression of Q_{sh}/Q_h against *s*, where Q_{sh}, Q_h and *s* are monthly mean global solar radiation (MJ·m^{-2}·d^{-1}), monthly mean extraterrestrial solar radiation (MJ·m^{-2}·d^{-1}) and monthly mean relative sunshine duration, respectively. Similarly, the coefficients of a, b and c in the horizontal diffuse solar radiation

model were determined from each month's observation (i.e. January, February, etc.), generating a set of coefficients for each month. The IDW interpolation method was then used to derive the spatial distribution of calibrated coefficients.

2.2.3.2 Radiation on the inclined surface

Global solar radiation on an inclined surface was calculated as the sum of direct, diffuse and reflected radiation from all sectors. This process was repeated for each grid cell in the DEM, thus producing an insolation map. Global solar radiation based on a DEM can be expressed as:

$$Q_{\beta w} = B_{\beta w} + D_{\beta w} + R_{\beta w}, \quad (6)$$

where $Q_{\beta w}$ is total solar radiation for rugged terrain. The direct, diffuse, and reflected solar radiation components within rugged terrain are $B_{\beta w}$, $D_{\beta w}$, and $R_{\beta w}$, respectively.

Similar to the clear-sky conditions on a horizontal surface, direct transmittance K_b was used to solve the atmospheric attenuation of direct radiation on a rough surface (Liu et al., 2012). Direct irradiance on the inclined surface can be expressed as:

$$B_{\beta w} = \frac{Q_{sw}}{Q_{sh}} \times B_h, \quad (7)$$

where Q_{sw} is slope extraterrestrial solar radiation.

In general, diffuse radiation coming from the sky is anisotropic. However, the calculation of anisotropy on a slope is complex and challenging (Dubayah and Rich, 1995). To simplify the calculation, diffuse radiation is divided into two parts: (1) one is from solar illumination direction; and (2) another is from isotropic modeling. The diffuse radiation is given by Zeng et al. (2008) as:

$$D_{\beta w} = D_h \times [K_b \times Q_{sw}/Q_{sh} + V \times (1 - K_b)] \quad (8)$$

When $k_b \to 0$, the sky is overcast and radiation is calculated from the isotropic model; when $k_b \to 1$, radiation is primarily from direct beam radiation. V is the sky view factor, which is associated with each grid cell. The detailed calculation process of V is illustrated in the supplementary information.

Radiation that is reflected from nearby surfaces (e.g., mountains) is a function of albedo, the sky view factor and horizontal solar radiation. The sky view factor is defined by the proportion of unobstructed sky over a horizontal surface such that $V = 0$ if the view of the sky is completely obstructed at a given location (Fu and Rich, 2002). Reflected radiation from nearby surfaces is calculated as:

$$R_{\beta w} = Q_h \times \rho \times (1 - V), \quad (9)$$

where $R_{\beta w}$ is radiation reflected by surrounding cells, ρ is surface albedo, in which ρ was determined as the ratio of reflected to incident solar radiation at the surface.

The algorithm for solar radiation over rugged terrain is calculated per pixel using iterative calculations for sunshine duration and sky view factor. Currently, the front-end JavaScript programming and backend GIS functions are not powerful enough in cloud-computing platforms to implement intensive and iterative algorithms (Gorelick et al., 2017b). To quickly obtain daily extraterrestrial solar radiation data over the vast area of the Loess Plateau (*ca*. 1×10^6 km^2), a parallel extraterrestrial solar radiation algorithm on a local HPC environment was developed using Python Multiprocessing and GDAL package for parallel processing. First, the multi-band image was split into tiles equaling 90% of the HPC cores. After running the algorithm, the Mosaic tool was used to combine the resulting tiles into complete and seamless time series of extraterrestrial solar radiation images of the Loess Plateau and sky view factor images.

2.2.4 Spatial and temporal MODIS albedo gap-filling

Albedo is composed of direct and scattered radiation components. Therefore, actual clear-sky albedo can be calculated by a linear combination of direct and scattered albedo:

$$\partial = (1 - S) \times \partial_h + S \times \partial_b, \quad (10)$$

where ∂ is land surface broadband albedo, ∂_h is broadband black-sky albedo, ∂_b is broad white-sky albedo and S is a conversion coefficient.

The spatial resolution of MODIS albedo is 1000 m, which is much finer than the surface

albedo data resolution of 8 km obtained by NOAA/AVHRR data. As seen in Figure 2-3, the gap-filled product showed a similar spatial distribution for high and low extremes of surface albedo. Albedo of various land types exhibited differences (Figure 2-4). For example, very high albedo was observed in desert lands (up to 0.7; Figure 2-4) during periods with snow cover. Outside of snow-covered periods, cropland, desert, and grassland land types showed a relatively stable albedo, with the highest albedo for desert landscapes (0.41 ± 0.23), intermediate albedo for crops (0.14 ± 0.03) and grasslands (0.13 ± 0.03), and lowest for forests (0.10 ± 0.04).

The main problem encountered in applying GIS-based methods is missing values of MODIS albedo. We overcame this problem using spatio-temporal correlation of un-gapped MODIS albedo data, interpolating across gaps using the surrounding data and producing a seamless dataset. This gap-filling method was comprised of five steps: (1) retrieve albedo data from the original MODIS albedo product; (2) create a mask of missing albedo data; (3) use a spatial neighborhood interpolation method to fill the missing data; (4) stack the yearly albedo image collection into the time series; (5) apply the Whittaker algorithm to smooth the time-series. This method was used to fill missing values without modifying existing values.

The Whittaker algorithm is based on penalized least squares, proposed by Whittaker 100 years ago (cited from Eilers, 2003). The Whittaker smoother has many advantages: it is extremely fast, much faster than the Savitzky–Golay filter in preliminary tests; it handles missing values efficiently; and it allows for full control over smoothness parameters (Eilers, 2003):

$$(W + \lambda D_d' D_d)\, z = W\, y \tag{11}$$

The missing elements of y are set to zero, and the diagonal elements of weight matrix W are set as zero for missing data and 1 for other values. At each missing point where y is zero, z was smoothed using Equation (11). D is a matrix such that $Dz = \Delta z$, and the subscript d represents the order of differences.

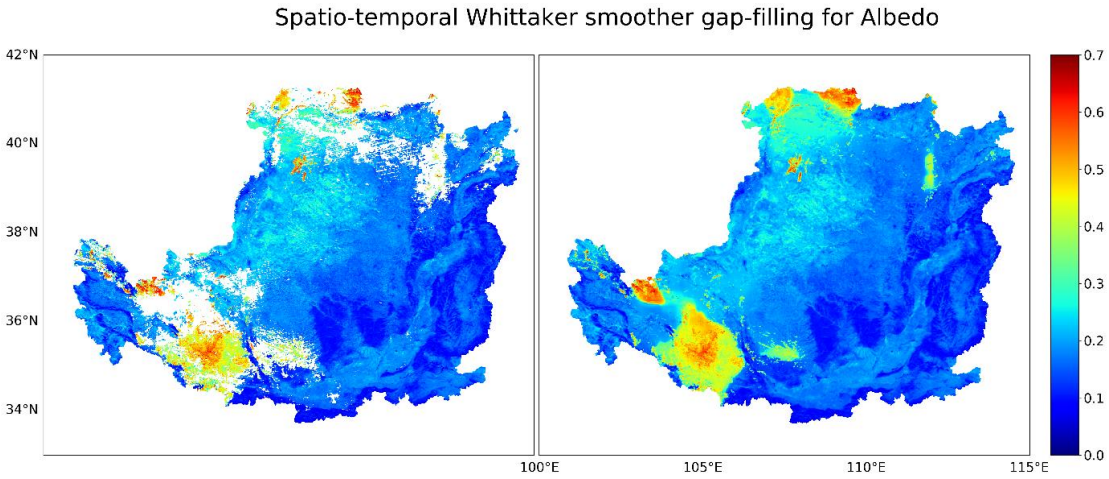


Figure 2-3 Albedo map of Loess Plateau at 1 January 2011 shown as an example of gap filling. Left panel shows missing values (white) in the northern and western regions of the plateau. Right panel shows Whittaker smoother gap-filled albedo map.

Figure 2-4 Variation of daily albedo for different land types. Missing values in raw albedo

images were filled by spatio-temporal gap-filling method. Those gap values in the curves were fitted by the Whittaker smoother method, with $\lambda = 20$, iterative=3.

2.2.5 Model evaluation

To evaluate the accuracy of the gap-fill predictions, Random Knockout Validation method (Gerber et al., 2018) was applied during one year (2011) to a rectangle areas that had fewer than 50% missing observations in the original data. 10 points were selected, representing 50%, 60%, 70%, 80% and 90% of the available original observations. 10 locations were randomly chosen from a validation area (Figure 2-5), and 10 temporal observations were randomly removed from each of these 10 time series (one time series per location). These missing values that were removed from the data were then filled using the spatial and temporal gap-filling algorithm described above, and then gap-filled values were compared to the originally removed observations. Figure 2-5 shows that gap-filled values at the 10 randomly chosen observation sites were significantly correlated with observed data resulting in an $R^2 = 0.962$ and an RMSE = 0.005 at daily timescales.

Validating and assessing the overall accuracy of STMSR based on only a few stations is inadequate for such a large region as the Loess Plateau. A comparison of STMSR estimates against those of other independent solar radiation products over China can provide an alternative approach for a regional evaluation of model performance. The comparisons include three key steps. First, daily SSR values were integrated from hourly values. Second, one thousand locations were randomly created across the extent of the Loess Plateau for comparison. Third, daily SSR values of the points in different datasets were retrieved from those 1000 locations using the point sampling method. This procedure was performed at the same 1000 locations for STMSR and SSR, where the RMSE between the two models was evaluated across the entire Loess Plateau.

The performance of STMSR was also compared with two other GIS tools, Solar Analyst (SA) in ArcGIS and r.sun in GRASS. In SA, the Points Solar Radiation tool was used to calculate time series of global solar radiation simulations at Yuzhong site because it contained

continuous global direct and diffuse solar radiation datasets. The diffuse proportion of the radiation parameter was set to 0.4 under clear-sky conditions, and the transmittivity parameter was set to a default value of 0.5. The r.sun program in GRASS cannot be used to simulate point radiation, but it is able to input parameters in raster format. Thus, a small patch of DEM around Yuzhong site (400 km^2) was clipped to simulate daily global solar radiation. Albedo was set to the default value of 0.2, and the linke turbidity coefficient was set as an annual average value of 1.9.

R^2 is a statistic that describes the goodness-of-fit for a model, while RMSE is used to measure the difference between values predicted by a model and those which were actually observed. We used these two statistical criteria to validate our model. Those two validation measurements were calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2} \tag{12}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}\left(O_i - \bar{O}\right)\left(P_i - \bar{P}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i - \bar{O}\right)^2}\sqrt{\sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2}} \right)^2 \tag{13}$$

where P_i and O_i are the predicted and observed daily surface solar radiation respectively, $\bar{O}$ and $\bar{P}$ are the mean daily surface solar radiation, i is the i^{th} sample, and n is the number of samples.

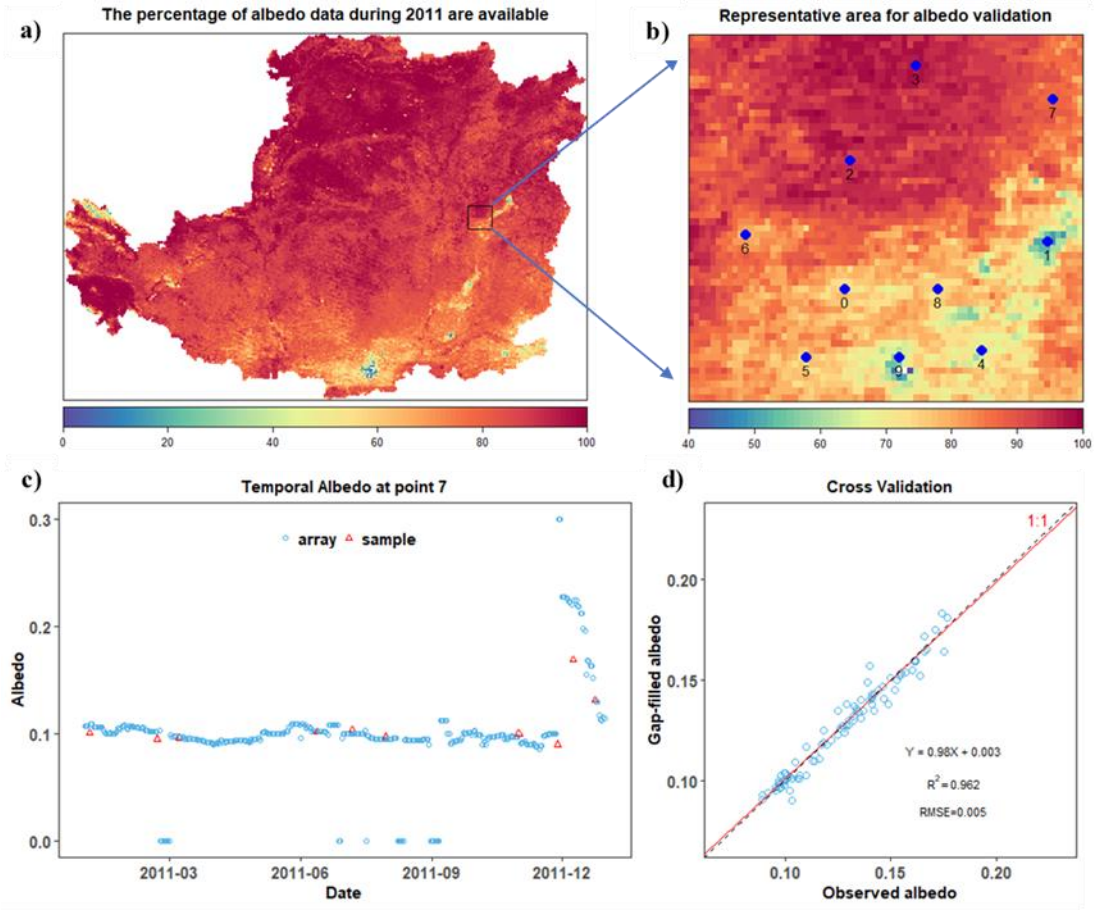


Figure 2-5 The percentage of albedo data during 2011 for the whole Loess Plateau (a), a representative validation area with 10 points (b), the temporal variation of daily albedo at point 7 with 10 randomly observed albedo (c) and cross validation for 100 samples during 2011 (d).

2.3 Results

2.3.1 Model validation in the Loess Plateau

Different timescales have Ångström coefficients (a_h and b_h) of varying magnitude. Previous studies have shown that a better fit between n/N and R_s/R_a can be obtained using monthly data than from yearly data. In this study, coefficients used for horizontal global and direct-beam radiation models were obtained from monthly in-situ radiation data, which were obtained from a previous study (Zeng et al., 2005).

To validate the performance of our model, observed data from ten solar radiation stations were used (Figure 2-1), of which one station (the Yuzhong station) included measurements of

both direct and diffuse solar radiation. Model performance for simulation of monthly solar radiation was evaluated for the period 2005–2009. In terms of global solar radiation, simulated monthly global solar radiation matched well with observations (Figure 2-6). Figure 2-6 shows that: (1) our model simulations at the 10 observation sites were significantly correlated with observed global radiation, resulting in high R^2 ($R^2 \geq 0.9$) and low RMSE (RMSE $\leq$ 45 MJ m^{-2} $month^{-1}$); and (2) the slopes were within ±10% of the 1:1 line across all study locations. At these 10 observations sites, our model performed very well in the Loess Plateau and can be further used to generate highly accurate solar radiation estimates for mountainous locations with local calibration/validation data.

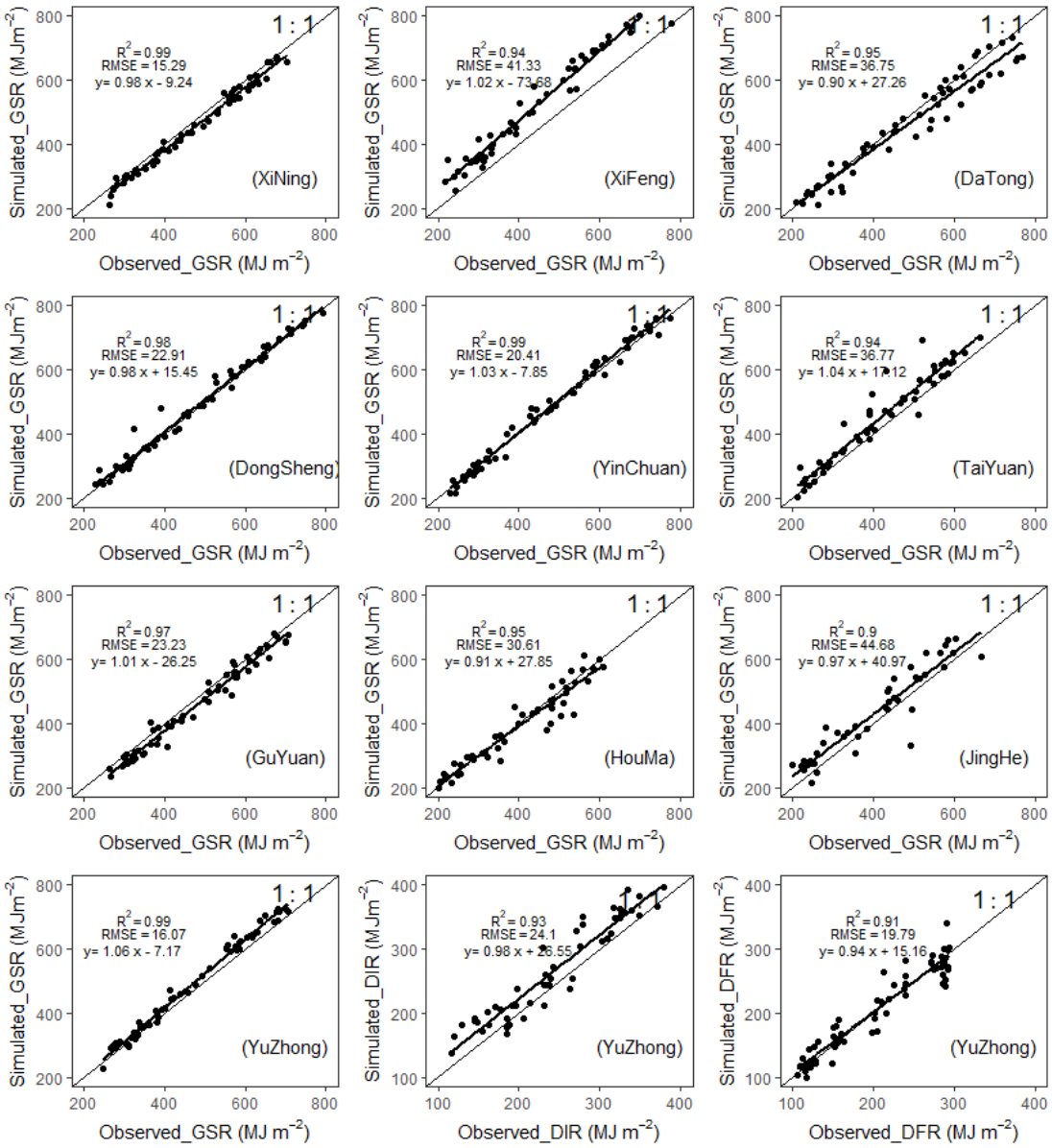


Figure 2-6 Comparison of annual observed and estimated (by mountain solar radiation model) monthly Global Solar Radiation (GSR) for 10 radiation sites on the Loess Plateau, China, during 2005 to 2009. Comparisons for direct radiation (DIR) and diffuse radiation (DFR) are shown only for YuZhong.

2.3.2 Comparison with other SSR and GSR products

Figure 2-7 illustrates the annual mountain solar radiation spatial map from STMSR, the SSR product, and the GLDAS net shortwave radiation product. In comparison to SSR, STMSR produced higher estimates of solar radiation in the drylands of the northwest Loess Plateau and

lower estimates in the mountains to the South (cf. Figure 2-7a, b). GLDAS net shortwave radiation values (lower left panel) showed little consistency with spatial patterns in STMSR or SSR and little association with topography (Figure 2-7). Maximum radiation was highest in STMSR (*ca*. 7000 $MJ \cdot m^{-2} \cdot d^{-1}$), intermediate in SSR (*ca*. 6500 $MJ \cdot m^{-2} \cdot d^{-1}$), and lowest for GLDAS (*ca*. 5500 $MJ \cdot m^{-2} \cdot d^{-1}$, Figure 2-7). STMSR similarly produced the lowest minimum radiation values (*ca*. 3500 $MJ \cdot m^{-2} \cdot d^{-1}$). Figure 2-8 shows performance comparisons between STMSR, SSR and GLDAS on a daily timescale, illustrating that solar radiation estimates from the current study were better than those from the other products. R^2 for STMSR (0.88) was better than that of the other two products (0.76–0.84), although R^2 for all was quite good (Figure 2-8). However, only two of the products (STMSR and SSR) showed a 1:1 response against observations (Figure 2-8). Overall, radiation estimates simulated by STMSR were slightly improved relative to SSR and greatly improved relative to GLDAS. We observed that STMSR–SSR RMSE (i.e., RMSE between two derived products, not observations) increased from north to south, indicating an increasing discrepancy between radiation products towards the South (Figure 2-9). Following this trend, RMSE was slightly higher in the Guanzhong Plain (in the southeastern Loess Plateau) than that in the mountains extending to the North along the eastern boundary of the Loess Plateau (cf. Figure 2-1 and Figure 2-9). RMSE was also large in the western-most region of the Loess Plateau, which has the highest elevations and steepest slopes of the Loess Plateau. Small discrepancies between STMSR and SSR in the central Loess Plateau suggest that both products produce reasonable radiation estimates. By contrast, larger discrepancies in the Guanzhong Plain and the western mountains suggest an improvement to radiation estimates by STMSR in these areas, thus STMSR can be further used to generate realistic solar radiation maps for mountain and valley regions.

Figure 2-10 shows a comparison of three algorithms against observed values at Yuzhong in 2009. The SA and r.sun algorithms clearly overestimated radiation in the middle of the year (March-October), and that SA underestimated radiation in winter months (November through the following February). In contrast, our STMSR model slightly overestimated observed

radiation March-October, but it closely predicted monthly global solar radiation across the remaining months. It should be noted that finer tuning of input parameters, such as direct transmittance and diffuse proportion in Solar Analyst or default atmospheric parameters in r.sun, might result in improved estimates from those products.

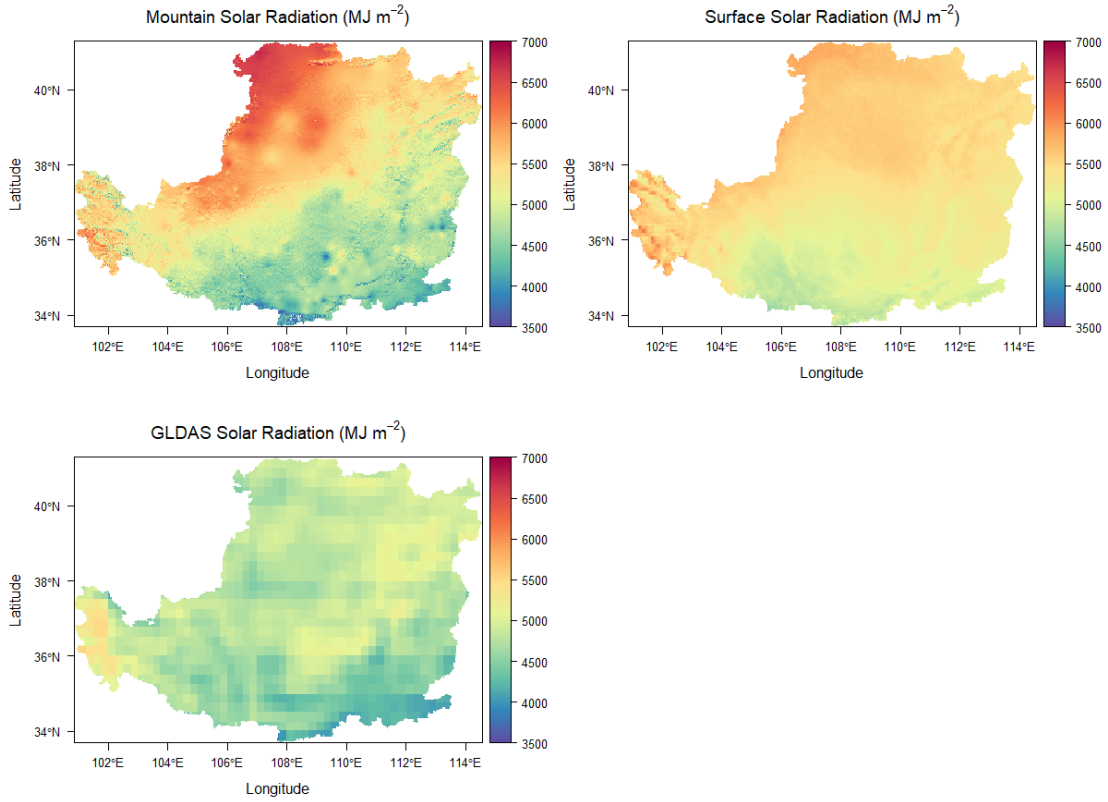


Figure 2-7 Spatial distributions of yearly solar radiation on the Loess Plateau in 2011 by mountain solar radiation produced by STMSR model, Surface Solar Radiation, and GLDAS.

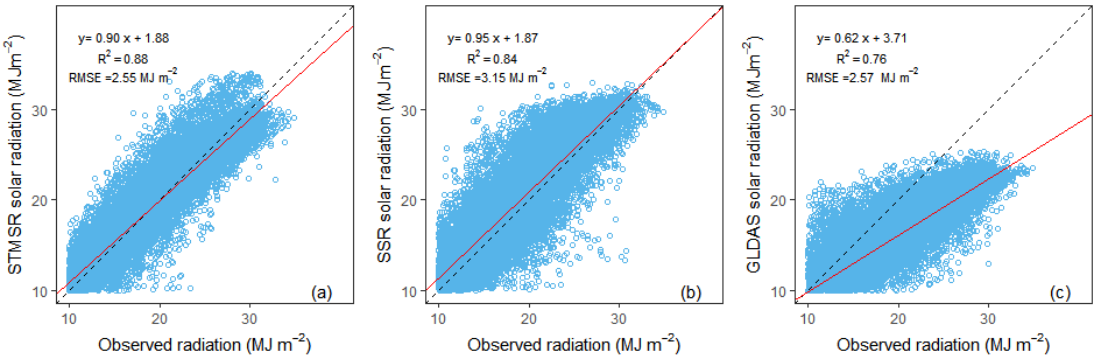


Figure 2-8 Summary statistics for estimated daily solar radiation produced by the Spatio-

temporal Mountain Solar Radiation (STMSR) model (a), the Surface Solar Radiation (SSR) model (b), and the Global Land Data Assimilation System (GLDAS) model (c) compared with observed data across 10 solar radiation stations in 2007-2013.

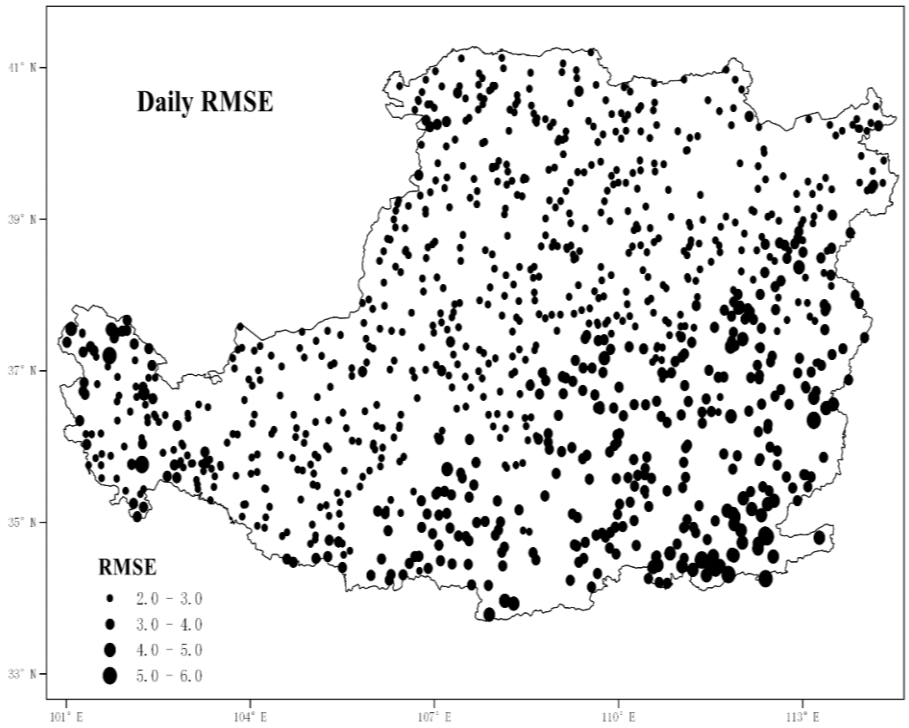


Figure 2-9 Spatial distributions of RMSE calculated between the daily solar radiation of the Spatio-temporal Mountain Solar Radiation (STMSR) model and Surface Solar Radiation (SSR) product at 1000 randomly selected points in 2011 over the Losses Plateau. Circle diameters correspond to the size of RMSE. RMSE units in the legend are MJ · m^{-2}.

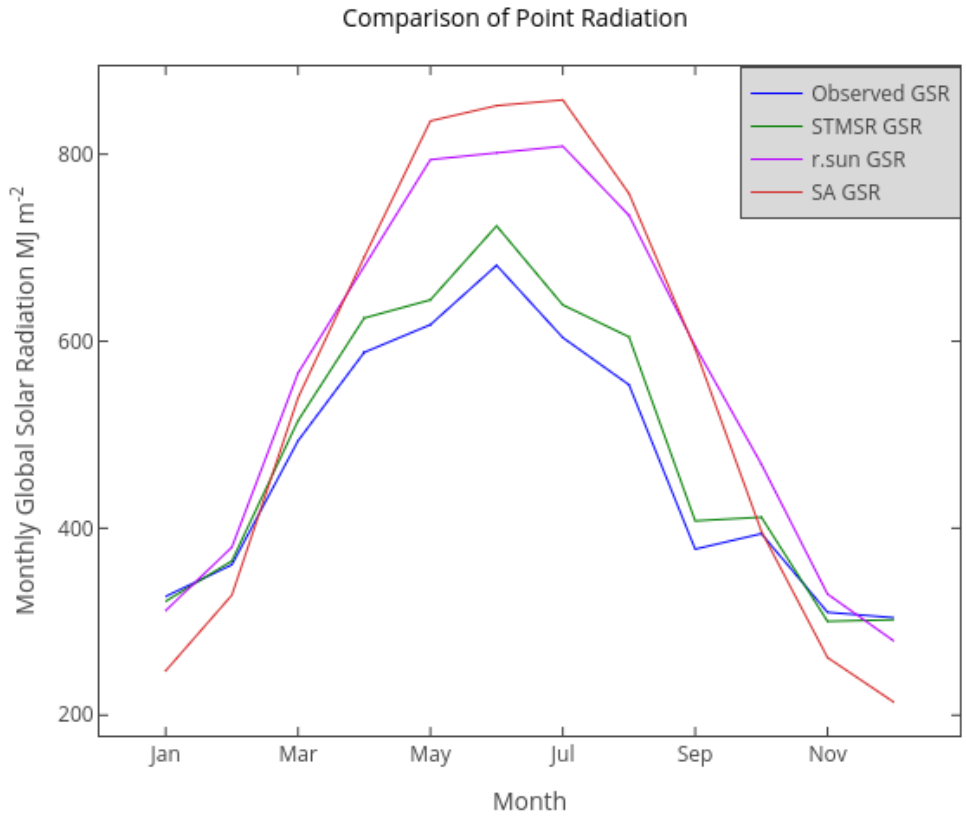


Figure 2-10 The comparison of different Global Solar Radiation (GSR) products with in situ observations at YuZhong in 2009. STMSR: Spatio-temporal Mountain Solar Radiation, SA: Solar Analyst and r.sun: radiation integrated in GRASS.

2.4 Discussion

We provide an online tool called "Spatial and Temporal Mountain Solar Radiation Modelling" (STMSR) as part of this study, available for use in complex terrain globally (https://geogismx.users.earthengine.app/view/stmsr) (Figure 2-11). In the left panel, users can define a time period along with a location by designating latitude and longitude, or by clicking on the map, and then the right panel will show a time series of point solar radiation, along with the three components of global solar radiation (direct, diffuse, reflected). STMSR can also export composited images of astronomical solar radiation (i.e., the radiation which would be incident at the planet's surface in the absence of an atmosphere) and global solar radiation components, as shown in Figure 2-12.

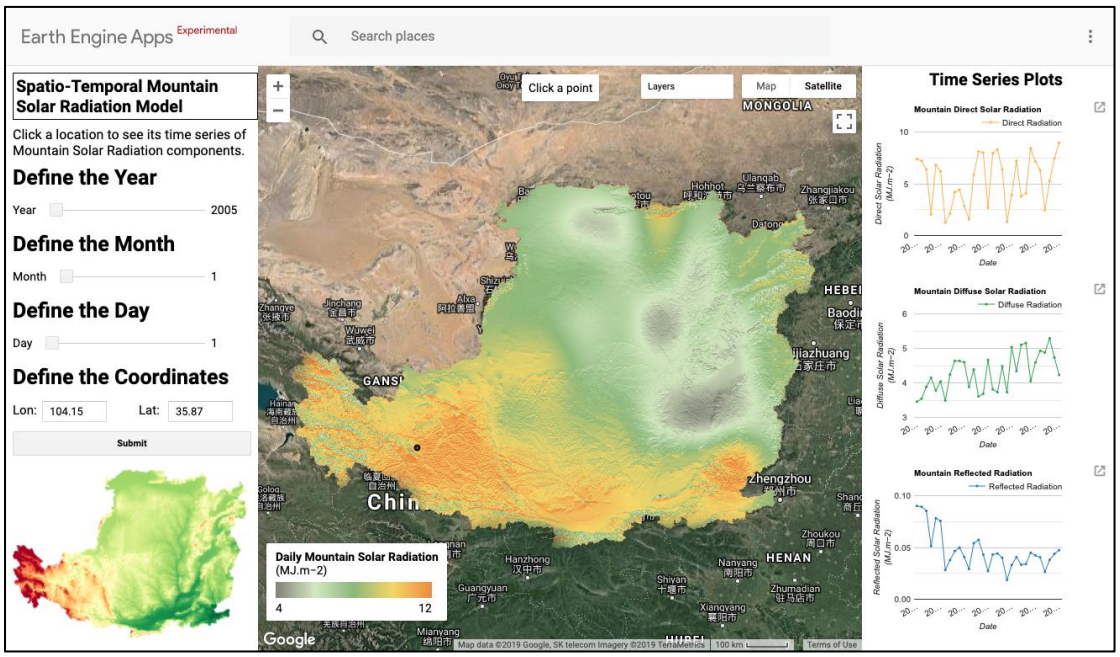


Figure 2-11 The application interface for the mountain solar radiation model on the Google Earth Engine APP Platform.

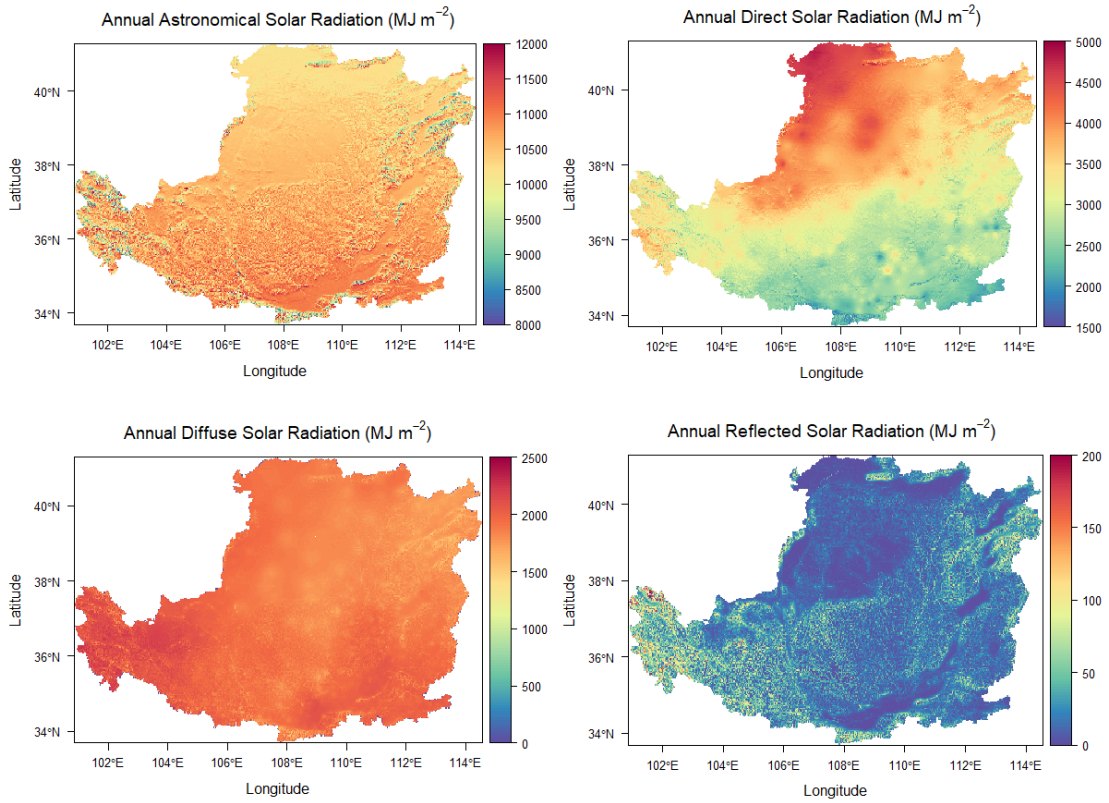


Figure 2-12 Spatial distribution of annual astronomical solar radiation, direct solar radiation, diffuse solar radiation, and reflected solar radiation in 2011 over the Loess Plateau.

The solar radiation estimated by our model performed better in the Loess Plateau than other SSR products and GLDAS, as quantified by R^2 and RMSE. This may be attributed to differences in satellite data sources, methods, and the scales of prediction. Bias in satellite-based models depends on clear sky index and solar zenith angle, together with atmospheric parameters, such as aerosol, ozone, precipitable water. Polar-orbiting satellites such as MODIS only measure instantaneous values, which are then extrapolated to daily solar radiation values using a sinusoidal function. This approximation is likely to incur a larger error than that provided by geostationary satellites like MTSAT (Qin et al., 2015; Roupioz et al., 2016). Even if MODIS and MTSAT measurements were integrated together to improve satellite-based solar radiation estimates, parameterization of the topographic correction remains overly simple. However, the time delay between atmospheric parameters derived from MODIS and actual

cloud variation can lead to significant errors. The spatial resolution of our mountain solar radiation estimates (about 1 km) is much finer than that of GLDAS net shortwave radiation data (0.25°) and SSR (5 km). The coarse spatial resolution of these two radiation products resulted in larger mean error relative to in-situ measurements than for STMSR.

Our open-source GIS-based model STMSR has both advantages and disadvantages. As described in the methods, one bottleneck was that GIS functions on the backend of cloud-computing are not powerful enough to implement iterative algorithms. In this study, some of those calculations were performed beforehand (e.g., duration of possible sunshine on slopes, the sky view factor). Then, they were uploaded to cloud storage for users to access widely and for incorporation in a cloud-based library of solar radiation models to decrease processing times dramatically. Another potential difficulty was that our model could be set up only when the coefficients of the sunshine-based model were available. However, this is not a problem because many solar radiation sites worldwide make available calibrated and accurate local direct and diffuse coefficients for sunshine-based models (Liu et al., 2009; Trnka et al., 2005). In such a case, the empirical parameterization scheme used in our model proved to be an economical and practical method for estimating actual solar radiation from sunshine hours under the influence of cloud cover. By contrast, satellite-based methods provide an advantage for retrieving atmospheric parameters from ungauged areas. Zhang et al. (2015) used two atmosphere products from MODIS, aerosol optical depth (AOD) and precipitable water (PW), as input parameters for solar radiation modeling to decrease atmospheric estimation errors. Concerning the rapidly rising array of satellite products becoming available, integration with more atmospheric products would be an important asset for future research.

As shown on Figure 2-3, the available daily MODIS albedo can be an issue in some regions for GIS-based solar radiation models due to data scarcity. Roupioz et al. (2016) chose to use the 8-day composite MODIS albedo product for the daily solar radiation modelling. However, this 8-day resolution is too coarse for investigating rapid changes in albedo over the Loess Plateau. We overcame this problem by developing a spatial and temporal gap-filling algorithm

to provide a seamless daily albedo dataset for estimating variations in solar radiation. This seamless dataset made possible the quick estimation of albedo over snowy landscapes, also providing further capabilities such as smoothing of other ecological indices and extracting phenological characteristics from data types such as NDVI, EVI or LAI (Pan et al., 2017). However, care must be taken when selecting the smoothing parameter (lambda) in the Whittaker algorithm, which is very sensitive to this parameter. In this study, lambda was determined by trial and error to be equal to 20, but further research is needed to evaluate the relationship between lambda versus kurtosis, mean and variance.

We found divergence between STMSR and SSR in the southern and western Loess Plateau (Figure 2-9). Large uncertainty in the southern portion of the Loess Plateau could be due complicated cloud distribution that reduces the accuracy of cloud parameter estimates, potentially leading to substantial errors in SSR estimation (Tang et al., 2016). By contrast, uncertainty in the western Loess Plateau was likely related to shading and surface inclination effects, both of the surface itself or in the adjacent terrain (Liu et al., 2012). Furthermore, spatial interpolation of regression coefficients across cloudy or mountainous regions can still be problematic (Liu, 2017) even though interpolation can be valid across some regions where the atmospheric turbidity is similar (e.g., across the central Loess Plateau, Figure 2-9). To overcome uncertainty due to interpolation errors, improved spatial and temporal interpolation of complex calibrated coefficients and sunshine hours in future studies could be achieved through integration of a Geographical and Temporal Weighted Regression (GTWR) (Fotheringham et al., 2015). Many factors can affect RMSE between radiation products, including interpolation of Ångström model coefficients, spatial variability in elevation, water vapor content, and other climate characteristics (Liu, 2017), reflective features of the surface, cloud contamination, aerosols, and atmospheric water vapor (Stocker, 2014).

Since the “Grain-for-Green” program has been implemented, large areas of re-vegetated land are now present in southern and eastern parts of the Loess Plateau (Zhang et al., 2018). Based on surface solar radiation theory, land use/land cover (LULC) can change

outgoing/reflected shortwave radiation and absorbed shortwave radiation by changing land surface albedo. However, it remains unclear whether LULC can change the incoming shortwave radiation reaching the land surface. LULC changes can be estimated by mean vegetation cover during the growing season, where vegetation cover is estimated by normalized difference vegetation index (NDVI). Future research could focus on changes in vegetation cover during the growing season for exploring the impacts of LULC changes on solar radiation.

2.5 Conclusion

We developed an improved GIS-based solar radiation model (STMSR, the spatial and temporal mountainous solar radiation model) that allows for treatment of high spatial and temporal variations in albedo, surrouding terrain shading and cloud cover for monitoring daily solar radiation at large scale. By comparison with other well-known GIS-based solar radiation models such as Solar Analyst in ArcGIS and r.sun in GRASS, our STMSR model showed better performance. The resulting estimates of global, direct, and diffuse solar radiation were validated with high estimation accuracy against the measured solar radiation data from 10 observation stations across Loess Plateau. Compared with other high-resolution solar radiation datasets, the global solar radiation presented in this paper has higher accuracy of daily solar radiation estimates over the Loess Plateau than other methods, generating higher R^2 and RMSE. Our STMSR model also has the potential to be applied globally for distributed modelling applications across a variety of landscapes.

Chapter 3. Creating new near-surface air temperature datasets to understand elevation-dependent warming in the Tibetan Plateau

This chapter is based on the following manuscript:

Zhang, M., Wang, B., Cleverly, J., Liu, D. L., Feng, P., Zhang, H., ... & Yu, Q. (2020). Creating new near-surface air temperature datasets to understand elevation-dependent warming in the Tibetan Plateau. Remote Sensing, 12(11), 1722.

Abstract

The Tibetan Plateau has been undergoing accelerated warming over recent decades and is considered as an early warning sign for broader global warming. However, our understanding of warming rates with elevation in complex mountain regions is incomplete. The most serious concern is the lack of high-quality near-surface air temperature (Tair) datasets in these areas. To address this knowledge gap, we developed an automated mapping framework for the estimation of seamless daily minimum and maximum Land Surface Temperature (LST) for the Tibetan Plateau from the existing MODIS LST products for a long period of time (2002-present). Specific machine learning methods were developed and linked with target-oriented validation and then applied to convert LST to Tair. Spatial variables in retrieving Tair, such as solar radiation and vegetation indices, were used in estimation of Tair, whereas MODIS LST products were mainly focused on temporal variation in surface air temperature. We validated our products using independent Tair products, revealing more reliable estimates on Tair, and the R^2 and RMSE at monthly scales generally fall in the range of 0.9-0.95 and 1-2°C. Using these continuous and consistent Tair datasets, we found a warming trend in the elevation range between 2000 m and 3000 m, whereas the summit above 6000 m exhibited a cooling trend. These datasets, findings, and methodology we developed contribute to global studies on accelerated warming.

Key words: *Near-surface air temperature; MODIS LST; machine learning; Tibetan Plateau*

3.1 Introduction

The Tibetan Plateau (TP) is named the "the third pole of the Earth", the highest and largest plateau globally (Qiu, 2008). The TP exerts profound dynamical and thermal influences on the regional and global climate (Duan et al., 2012; Manabe and Terpstra, 1974). For global warming, TP is considered as an early warning sign. Over the period of 1984-2009, TP has undergone serious warming, with a warming rate of 0.46°C decade^{-1}, which is almost 1.5 times the rate of global warming (Kang et al., 2010; Kuang and Jiao, 2016). Accelerated warming on the TP has intensified permafrost degradation, snow melt and glacier retreat (Yang et al., 2014). Presently, the status of TP warming is evaluated through the analysis of Tair at meteorological stations. However, most meteorological stations are located in the eastern TP below 3800m. Because of sparse high-elevation meteorological observations in central and northwest of TP, there is a possibility that we may not capture the greatest warming rate over some regions of the TP (Dimri et al., 2019). In addition to limited coverage by in-situ measurements, Tair at TP suffers from extreme local variability due to factors such as topography and exposure (Pepin et al., 2015). Therefore, improved Tair estimations by developing high-resolution products considering rugged terrain over the TP is a crucial step for understanding the accelerated warming in the TP.

Remotely sensed Land Surface Temperature (LST) is a crucial parameter in the modelling of surface energy balance at regional and global scales (Anderson et al., 2008; Fu et al., 2019; Mallick et al., 2014). LST also provides the possibility of getting high spatial-temporal daily Tair datasets (Tair). Since the late 1980s, a variety of long-time series LST products from Moderate Resolution Imaging Spectroradiometer (MODIS), Advanced Along-Track Scanning Radiometer (AATSR), Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), Multi-Functional Transport Satellite (MTSAT), and Geostationary Operational Environment Satellite have been published (Ouyang et al., 2017). MODIS Terra/Aqua sensors provide close temporal proximity of overpasses, with four LST datasets per day. Many studies have focused on using various combinations of the four MODIS LST datasets to estimate Tair (Noi et al., 2017; Yang et al., 2017; Zhang et al., 2016). Studies have also investigated how to combine the four LST datasets for creating composite daily minimum and maximum LST values that supplement the existing Terra/Aqua LST products and reduce areas of missing data. Crosson et al. (2012) increased data coverage of MODIS LST in the United States by 24% and 30% for daily minimum and maximum LST, respectively. Li et al. (2018a) developed a hybrid gap-filling method that merged dataset from existing LST products to fill gaps while integrating

with a spatio-temporal gap-filling method to fill the rest of MODIS daily LST gaps, finally creating LSTs dataset over the urban and surrounding areas of United States. However, to the best of our knowledge, no attempt has been made to explore an automated mapping framework for the estimation of daily seamless minimum and maximum LST from the four MODIS LST products.

LST is not equivalent to Tair and their relationship is complex from a theoretical and empirical perspective (Yang et al., 2017). Hence, it is difficult to estimate surface air temperature solely using LST, and additional auxiliary factors are used to estimate the Tair in mountainous regions using three representative methods. (1) A semi-empirical method which builds a linear relationship between MODIS LST and a vegetation index, such that Tair can be extrapolated by allowing the regression line to intersect with the vegetation index of full cover (Stisen et al., 2007; Zhu et al., 2013). (2) A spatio-temporal regression method has been used, such as a geographically and temporally weighted regression (GWTR) or a regression-kriging method, both of which consider the relationship between Tair and other variables such as MODIS LST and topographical layers (Kilibarda et al., 2014; Metz et al., 2017). (3) Machine learning models predict Tair from multiple data sources including LST, whilst considering spatio-temporal autocorrelation of Tair (Zhang et al., 2016; Zhu et al., 2019). In general, each method has been proven to be successful in estimating Tair, but they still have shortcomings in large-scale complex mountainous area with limited weather stations, especially in TP. For example, the Temperature-Vegetation Index (TVX) method is not appropriate for estimating Tair in regions with low vegetation cover (Yoo et al., 2018). Statistical models fail to capture the nonlinear behaviour of the climate system in mountainous area. Machine learning models such as Random Forest, Cubist and Support Vector Machine (SVM), have proven to be flexible in areas with complex terrain like TP for estimating Tair from LST and additional variables (Yoo et al., 2018). But they require more datasets for model training. The problem is that training datasets are usually insufficient in such complex mountainous areas. In addition, machine learning models often fail to capture the extreme low and high values of Tair (Kalra and Ahmad, 2009; Leihy et al., 2018). Furthermore, the estimated performance of machine learning models has a risk of spatio-temporal over-fitting, which partly depends on different validation strategies (Meyer et al., 2018). Thus, further investigation is required to identify the accuracy and performance of machine learning models when lacking enough field observations in complex mountainous areas of TP.

Other environmental datasets, such as vegetation indices, snow cover, albedo, soil type and water bodies, are highly related to Tair. LST with those auxiliary information enables the

estimation of Tair within mountainous areas using machine learning models. It is well known that the variation of incoming solar radiation has a strong relationship with the spatial temporal dynamic of Tair (Bristow and Campbell, 1984). In previous studies, solar zenith and sunshine duration were used as substitutes of mountainous solar radiation for Tair estimation (Zhang et al., 2016). In this study, we will incorporate truly mountainous solar radiation datasets as one of the covariates for Tair estimation (Zhang et al., 2020). We assumed that solar radiation and biophysical factors would be related to spatial variability in Tair, whereas we predicted that MODIS LST would be more strongly related to temporal variation in Tair.

The objectives of this study were to (1) create seamless 1000 m daily MODIS LST datasets using a hybrid method; (2) predict Tair using LST and remotely sensed indices with machine learning; (3) compare the performance of different machine learning methods for estimating maximum, minimum and mean air temperature; accordingly, and (4) explore elevation-dependent warming over the TP using decadal temperature datasets.

3.2 Materials and Methods

3.2.1 Study area and all climate data

The TP is located at 26-40 N and 73-105 E degree. It has irregular topography with elevation varying between approximately 498 and 7198 m a.s.l. (above sea level) and generally increasing from northwest to southeast (Figure 3-1). As elevation increases, the landscape transitions from forests to alpine grassland and then bare rock, and finally to snow and ice (Pepin et al., 2019). The highest Himalayan mountains are on the southern edge of TP, while the Kunlun Mountains are another high mountain chain on the northwest boundary. The headwater areas of major rivers in Asia lie in the south-eastern part of TP (Hengduan Mountains). Typical alpine permafrost lies in Bayan Har Mountains. Qaidam Basin is the largest terrestrial basin of the TP.

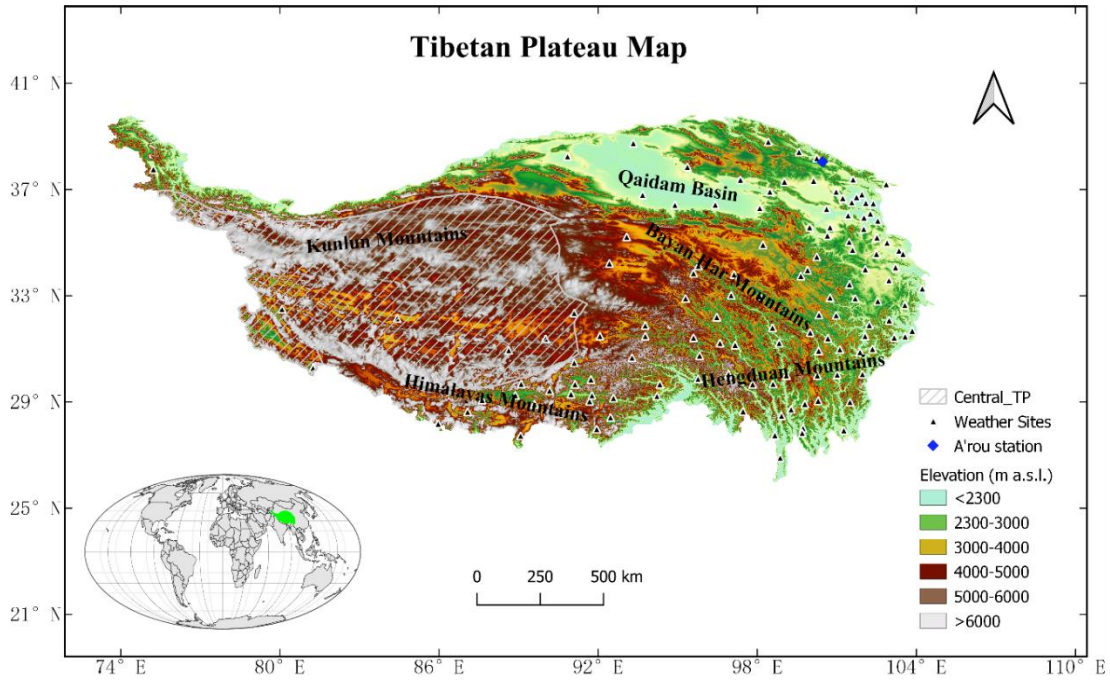


Figure 3-1 Location of Tibetan Plateau, distribution of 130 weather stations and A'rou station

In this study, daily observations of Tmax, Tmin, Tmean, and sunshine duration (2000–2016) from 130 available China Meteorological Administration (CMA) stations were used, the altitude of those stations ranges from 1600 to 4800 m a.s.l. To keep consistency between MODIS LST pixels and ground observations, a relatively homogeneous LST validation site named A'rou was chosen, which was located on the northeast edge of the TP with an elevation of 3032 m a.s.l. The in-situ LST was derived from the upwelling and downwelling longwave radiation fluxes from A'rou station using a radiation transfer equation. As different emissivity sources provide different accuracies in LST calculation, in this study we used ASTER-derived emissivity, derived from clear-sky pixels of ASTER images from 2000 to 2008 covering the study area. The monthly mean Tair observation data over the TP in 1981-2010 used in this study as reference data are from the China Meteorological Data Service Centre (http://data.cma.cn/site/index. html). Another dataset is from the China Meteorological Forcing Dataset (CMFD) for the period of 1979-2018, developed through fusion of remote sensing products, reanalysing datasets and in-situ station data with a spatial resolution of 0.1° and a temporal resolution of three hours. Due to its continuous temporal coverage and consistent

quality, the CMFD is one of the most widely-used climate datasets for China (He et al., 2020). In addition, TerraClimate is a global gridded dataset of meteorological and water balance variables at 2.5 arc-minute resolution (4000 m) from 1958 to 2020 (Abatzoglou et al., 2018). TerraClimate updates on monthly time step, and available at https://climate.northwestknowledge.net/ TERRACLIMATE.

Table 3-1 Overview of datasets across the TP

Data Source	Temporal Resolution	Spatial Resolution
DEM	2000	30 m
Weather Sites	2003-2013, daily	
LST Site	2007-2011, 10-min	
MOD11A1/MYD11A1	2002-Current, daily	1000 m
MOD09GA	2000-2020, daily	1000 m
CMFD	1979-2018, 3-hour	0.1 degree
TerraClimate	1958-2018, monthly	0.025 degree

3.2.2 Methodology

Spatio-temporal patterns of Tair in mountainous areas were quite complex due to the influence of landscape-scale physiographic factors. To address these problems, we developed a modelling framework for Tair (Figure 3-2). The first step was that we apply a hybrid method (combine serval method, i.e., daily merging and spatio-temporal gap-filling) to create seamless remotely sensed LST datasets. The second step was to calculate a set of predictors including LST datasets, mountainous solar radiation, biophysical factors and topography indices for surface air temperature modelling over the TP from 2003 to 2013. The calculation of the above two steps were all conducted in the Google Earth Engine (GEE) environment. Those predictors were then integrated as explanatory variables for machine learning models, while the in-situ measurements were used as the response variables. To avoid the spatial-overfitting of machine learning models, target-oriented validation strategies were used. In the third step, the cross validation (CV) data was split into 10 folds using spatial ID and Year ID as splitting criterion to predict on unknown points in time and unknown locations. Then, the best model for each

month was used for final tuning with 10-fold Leave Location and Time Out-Cross Validation (LLTO-CV) allowing us to provide accurate monthly spatial maps of Tair over the TP. The final step was to evaluate the climate change in TP with monthly Tair datasets by comparing with other temperature products. More details were presented in the following subsections.

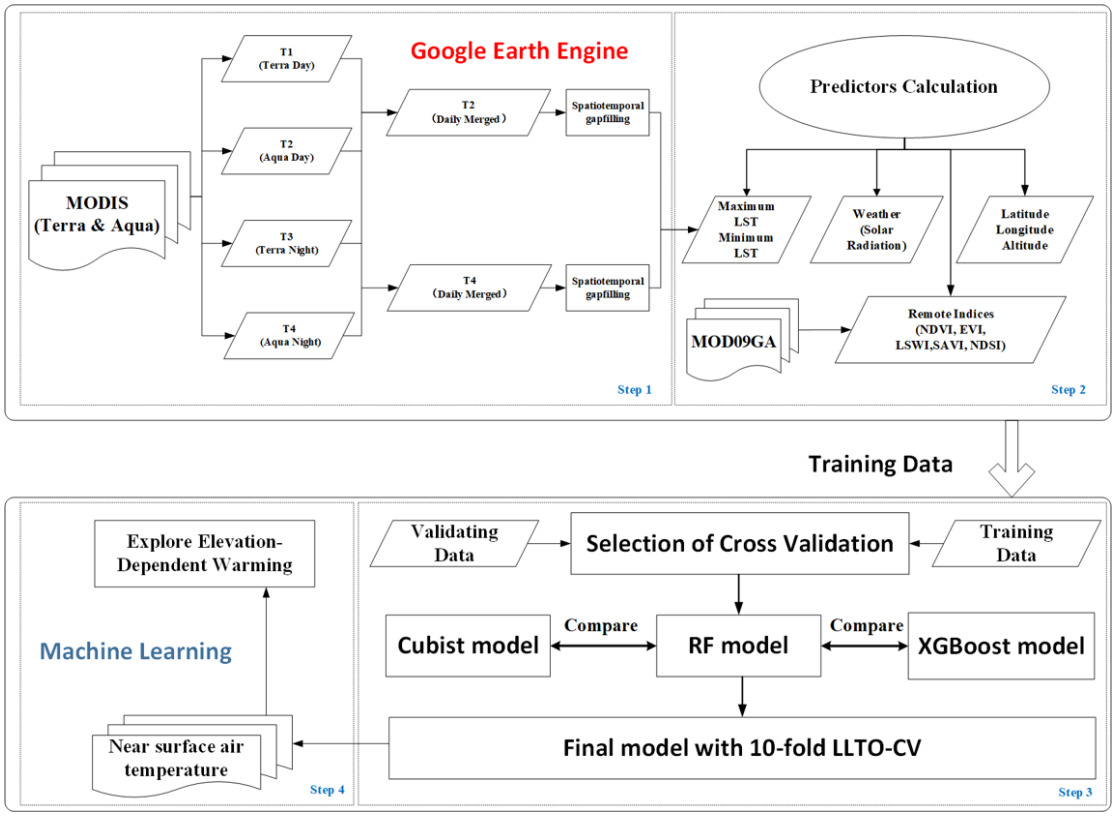


Figure 3-2 Flowchart of steps for calculation of near-surface temperature over TP

3.2.2.1 Step 1: Hybrid model to estimate daily seamless MODIS LST and validation

Development of globally complete spatial-temporal daily LST images still face many challenges. Considering the advantages of MODIS LST including 1000 m spatial resolution with high temporal resolution and the computing efficiency of different gap-filling methods, we used a three-step hybrid method to build daily LST. The first step was the daily merging method which involved using values from the other three times on the same day to fill the missing values for a given time. For example, we estimated T2 from T1, T3 and T4 and obtained four time series of T2 (LSTday) and then composited them to get the merged LST on daily basis. The details of this step were presented in Li et al. (2018a). The benefit of the daily merging method of four observations is that it increases the spatial and temporal extent of the

daily LST coverage. The second step was to use a spatio-temporal gap-filling method by estimating missing values with values of their neighbouring cells and days. Existing spatio-temporal gap-filling methods are all sensitive to parameter configurations. However, when applied at a global scale, a set of universal key parameters are difficult to select. In this study, we used a new spatio-temporal gap-filling algorithm instead of using traditional gap-filling packages or software as this method is computationally efficient and is promising for large scale applications. Especially, we chose 10000 m as the searching radius of bicubic interpolation and 30 days as the given window for moving average temporal interpolation. The third step was to use Whittaker smoother to remove the outliers introduced by spatio-temporal filled LSTs. We noted that the hybrid method depends heavily on daily merging LSTs. Therefore, the daily merging LSTs are treated as good LSTs and the remaining gaps left after the daily merging are filled by Whittaker smoother values.

Validation of land surface temperature used the radiation transfer equation below:

$$R_n = R_{si} - R_{so} + R_{li} - R_{lo} \quad (1)$$

where R_{si}, R_{so}, R_{li}, R_{lo} are incoming shortwave radiation ($w \cdot m^{-2}$), outgoing shortwave radiation ($w.m^{-2}$), incoming longwave radiation ($w.m^{-2}$), outgoing longwave radiation ($w \cdot m^{-2}$) respectively, R_n with the net radiation. In Equation (1), R_{lo} is a direct function of land surface temperature. For a surface with emissivity ε (unitless), the outgoing long-wave flux is composed of both reflected and emitted parts.

$$R_{lo} = (1 - \varepsilon) \times R_{li} + \varepsilon \times \sigma \times T_s^4 \quad (2)$$

where σ is Stefan-Boltzmann constant ($5.67 \times 10^{-8} \cdot W \cdot m^{-2} \cdot K^{-4}$), T_s is land surface temperature. If R_{li}, R_{lo} were obtained from radiometric network, the accuracy of the land surface temperature is dependent on the emissivity ε.

3.2.2.2 Step 2: Remotely sensed indices, DEM derivatives and mountainous solar radiation

Due to the limited data availability over TP, the model accuracy largely relies on the input datasets. If it is insufficiently trained, a robust correlation cannot be expected, that is, the more sufficient the samples, the better the spatial prediction accuracy (Li et al., 2018a). MODIS Terra/Aqua LST products (MOD11A1 and MYD11A1) were obtained from NASA (https://lpdaac.usgs.gov/products/myd11a1v006/,https://lpdaac.usgs.gov/products/mod11a1v0

06/). The MODIS Terra overpass time is around local time 10:30 AM (T1) in its descending mode and 10:30 PM (T2) in its ascending mode. The MODIS Aqua overpass time is around 1:30 PM (T3) in its ascending mode and 1:30 AM (T4) in its descending mode. Those time-variant biophysical factors such as Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Land Surface Water Index (LSWI), Normalized Difference Snow Index (NDSI), and Soil Adjusted Vegetation Index (SAVI) derived from MODIS surface reflectance bands, which has the same spatial and temporal resolution with LST, were incorporated as training inputs for machine learning models. This study also produced incoming mountainous solar radiation datasets as a predictor to estimate air temperature distribution in TP and provided a web app tool based on GEE for user access (https://geogismx.users.earthengine.app/view/tpmsr). In fact, incoming mountainous solar radiation in TP was the first time used in this study as previously such datasets were not existing and not used as a variable to estimate air temperature over TP. Therefore, with the availability of topographical variables and weather datasets and biophysical factors in TP, we had the potential to provide a more interoperable and rigorous way for estimating reliable Tair at high mountainous area.

3.2.2.3 Step 3: Regression models and target-oriented validation

In this study, we adopted three commonly used regression techniques to reproduce Tair. From the literature review, it was suggested that the differences in land surface properties, solar radiation, topography, and many other factors could influence the relationships between MODIS LST and Tair. Therefore, linear regression model is unlikely to be able to handle the complicated relationship between Tair and the abovementioned variables under different conditions. In contrast, advanced machine learning models, such as Random Forest, Cubist and XGBoost, can take account of the nonlinear and complicated relationship between the predictor and response variables in a mountainous study area like TP.

Random Forest (RF), also known as random decision forest, is an advanced ensemble machine learning technique which can be used to develop predictive models for both regression and classification purposes (Breiman, 2001). The ensemble technique is an algorithm that integrates outputs from multiple learning models to generate a better prediction. In the case of

RF, it achieves this goal through obtaining outputs from a whole forest of random decision trees. Decision trees are also a popular regression method, but they tend to overfit on training data and usually have high variance even if utilizing different training and testing sets from a same dataset (Olaru et al., 2003). Nevertheless, decision trees can be used as an underlying foundation in ensemble methods for producing more accurate predictions. The RF first creates an ensemble of decision trees through a process of bagging (bootstrap aggregating). Randomized subsets of the predictors are assigned to each tree to generate predictions. The average of the predictions from the ensemble of the trees is treated as the final outcome (Cutler et al., 2007). Thus, the RF succeeds in reducing the variance by creating a majority-votes model. In recent years, RF has been frequently used in remote sensing related research (Hashimoto et al., 2019; Moreno-Martínez et al., 2018).

eXtreme Gradient Boosting (XGBoost) is a scalable and efficient implementation of the gradient boosting framework (Friedman, 2001). It is also an ensemble technique that can build a final predictive model based on a larger number of underlying models. The most commonly used underlying model is a regression tree, different to RF. Another difference is that XGBoost repeatedly trains trees or the residuals of the previous predictors, while RF trains many independent trees and then average them. In the present study, we adopted DART (Dropouts meet multiple Additive Regression Trees), an ensemble model of boosted regression trees, which is capable of overcoming the issue of "over-specialization" (Rashmi and Gilad-Bachrach, 2015).

Cubist regression is a commercially rule-based regression method that was developed based on a combination of the ideas of Quinlan (1992). That is why it lacks algorithmic documentation. After Cubist regression was introduced into R by Kuhn, it has been widely used in remote sensing studies. Unlike CART-based regression trees (e.g., RF) that have a final model, Cubist produces rule-based multivariate regression models, which means that a set of rules is associated with sets of multivariate regression. Then, an actual prediction model will be chosen based on the rule that best fits the predictors. Since Cubist generates rule-based results, it is more straightforward and interpretable than RF. Cubist has much shorter run time than CART-based regression trees.

K-fold Cross Validation is popular to estimate the performance of the model with a view towards data that has not been used for model training. The validation dataset is randomly split into k folds during standard random k-fold Cross Validation. However, the problem of dependencies caused by the nature of spatial-temporal data was ignored, producing an over-

optimistic model performance because of spatial-temporal over-fitting (Gasch et al., 2015; Meyer et al., 2016; Meyer et al., 2018). To be more specific, many prediction models use auxiliary predictor variables which vary in space but not in time (e.g., elevation, location, and biophysical characters). However, those temporally static variables that focus on describing the spatial characteristics of the climate stations are prone to enable machine learning algorithms to disguise real relationships between predictors and responses and lead to spatial over-fitting. For example, the performance differences between K-fold random Cross Validation (lower RMSE) and Leave-Location-Out Cross Validation (higher RMSE) in the literature strongly suggest that spatio-temporal prediction models fail in the prediction beyond the location of training stations but can very well predict on the unknown time of the training stations.

As we aim to predict air temperature in unknown locations, we perform a target-oriented validation which validates the model with a view towards spatial mapping. To find this out, we repeatedly leave the complete time series of one or more data loggers out and use them as test data during CV. This study will use the following two steps to identify and avoid over-fitting.

1. To compare Machine Learning methods with different validation strategies using 10-fold Leave-Location-Time-Out (LLTO), Leave-Location-Out (LLO) and Leave-Time-Out (LTO).

2. Using the best fitting model with suitable validation strategies to estimate monthly Tair products based on 10-fold LLTO Cross-Validation.

3.2.2.4 Step 4: Creating near-surface air temperature products and elevation-dependent warming analysis

The controversy over the elevation dependent warming of TP mainly because we lack high-altitude meteorological data over TP (Li et al., 2020a). In this paper, to obtain more reliable Tair products for elevation dependent warming analysis, we produced 11 variables to build the actual non-linear relationship with response and tested three machine learning models with three different validation methods. According to the performance measurement of RMSE and R^2, RF was finally selected as the best model for Tair products generation. By comparing the accuracy of the three Tair products, the most accurate monthly mean air temperature during 2003 to 2013 was selected to analyse the temperature change over TP.

Due to the wide spatial domain of TP, temperature variations are inconsistent in different

regions. Therefore, in this study all the pixels in the three 1000 m elevation interval over TP were extracted and the time series of temperature changes in each elevation range were computed from the mean of the pixels. Specifically, we explore the relationship between temperature trends and elevations in 2000-3000 m, and 4000-5000 m and 6000-7000 m a.s.l., respectively. Among the analysis of temperature trends, the Seasonal Mann–Kendall statistical test and Seasonal Sen's slope test (Hirsch et al., 1982) were employed to test the significances of trend and the magnitude of trend in the seasonal mean temperatures.

3.3 Results

3.3.1 Evaluation of spatio-temporal composite LST

Figure 3-3 shows the percentages of all available days per year for which maximum LST and minimum LST before and after using daily merging method. Overall, the percentage of MODIS LST data availability over TP is over 80%, while after daily merging it is over 99%. For example, daily merging of the four observations increases most of LST data coverage by about 30% and 20% for LST day and LST night, respectively. In addition, a comparison of Figure 3-3(a) and Figure 3-3(c) shows that LST day values availability is intrinsically lower than that in LST night. For the central part of TP, the data coverage of observed LST day is around 70%, while the data coverage LST night availability is about 80%. For the eastern part of TP, available percentages for observed LST day and night are about 45% and 60%, respectively. After merging daily LSTs from the four overpasses, LST coverage can increase to over 99% in central TP, and about 80% in eastern TP. However, it still fails in filling the gaps in the boundary of southern TP (see in Figure 3-3(b) and Figure 3-3(d)). Therefore, an additional step is to use spatio-temporal gapfilling to fill the remaining missing values for the whole regions.

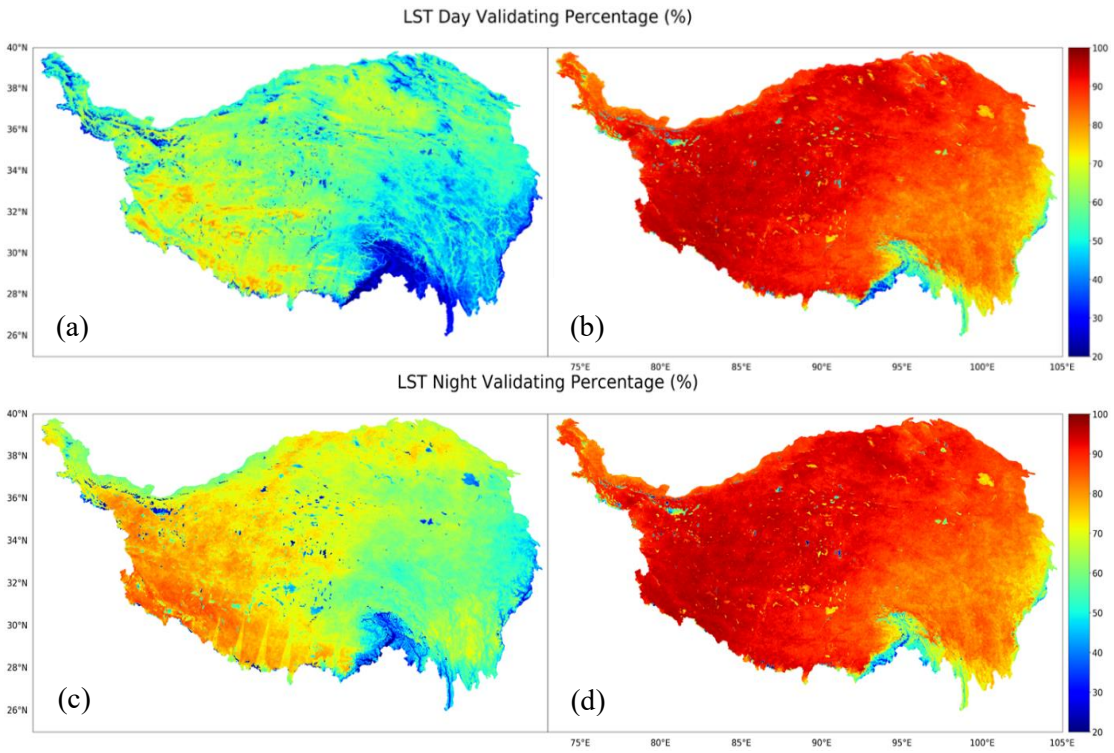


Figure 3-3 shows the prevalence of available data in the two pairs of maps. Figure 3-3(a) shows the percentage of days for the given year for which LST day (i.e. 1:30 pm on Aqua (T2)) values are available at each pixel of the TP domain. Figure 3-3(b) shows the percentage of daily merged T2 for the given year for which daily merged T2 values are available. Figure 3-3(c) shows the percentage of days for the given year for which LST night (i.e. 1:30 am on Aqua (T4)) values are available at each pixel of the TP domain. Figure 3-3(d) shows the percentage of daily merged T4 for the given year for which daily merged T4 values are available.

To evaluate the accuracy of MODIS spatio-temporal composite LST, the ground measurements comparison at A'rou station for 2008 was conducted for both maximum LST and minimum LST observations. The ground measurements at the nearest collection times matching with MODIS maximum and minimum LST were used. The annual comparisons are shown in the left panel of Figure 3-4. The scatterplots of the comparisons were given in the right panel of this figure. For the maximum LST comparison, the RMSE and R^2 over A'rou station were 5.44 °C and 0.76. For the minimum LST comparison, the RMSE and R^2 were 5.14 °C and 0.77. It should be noted that the minimum LST results were slightly better than for the maximum LST due to the stable weather conditions at night. From the annual results comparison, we can see that during the winter/spring when the freeze/thaw transition happens frequently, the

differences of maximum LST or minimum LST with in-situ measurements were greater than that of other seasons.

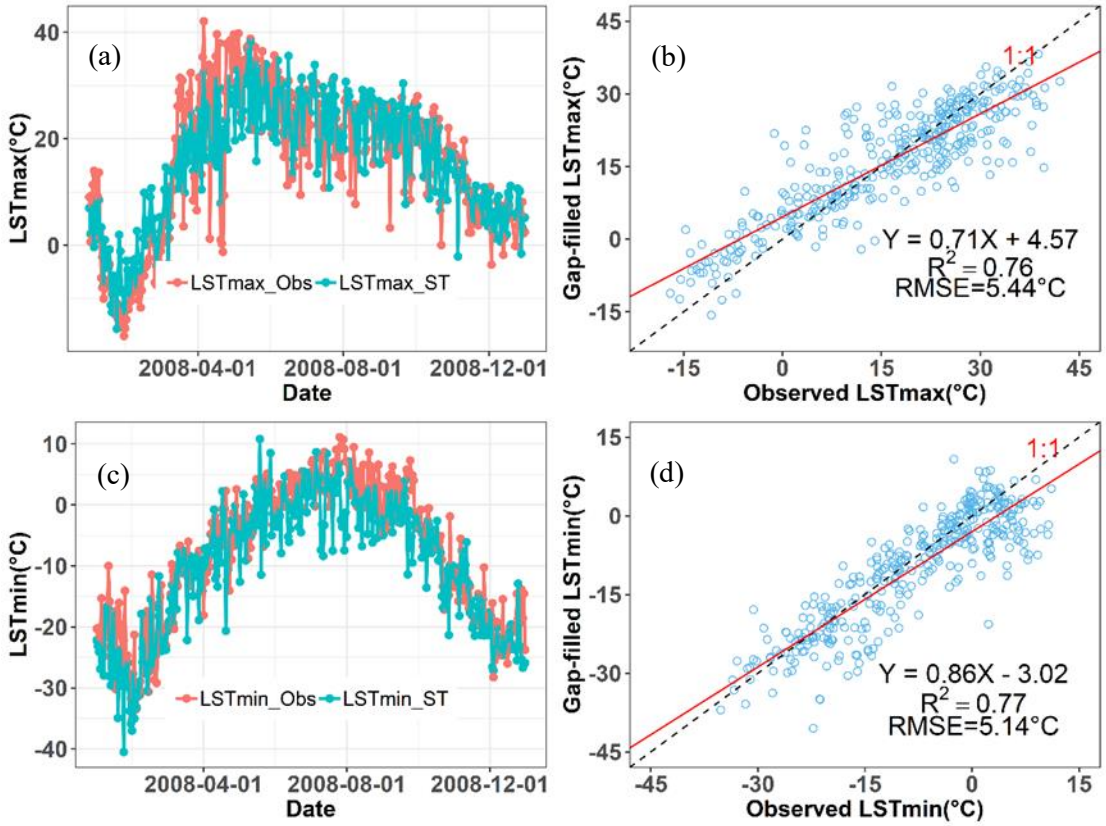
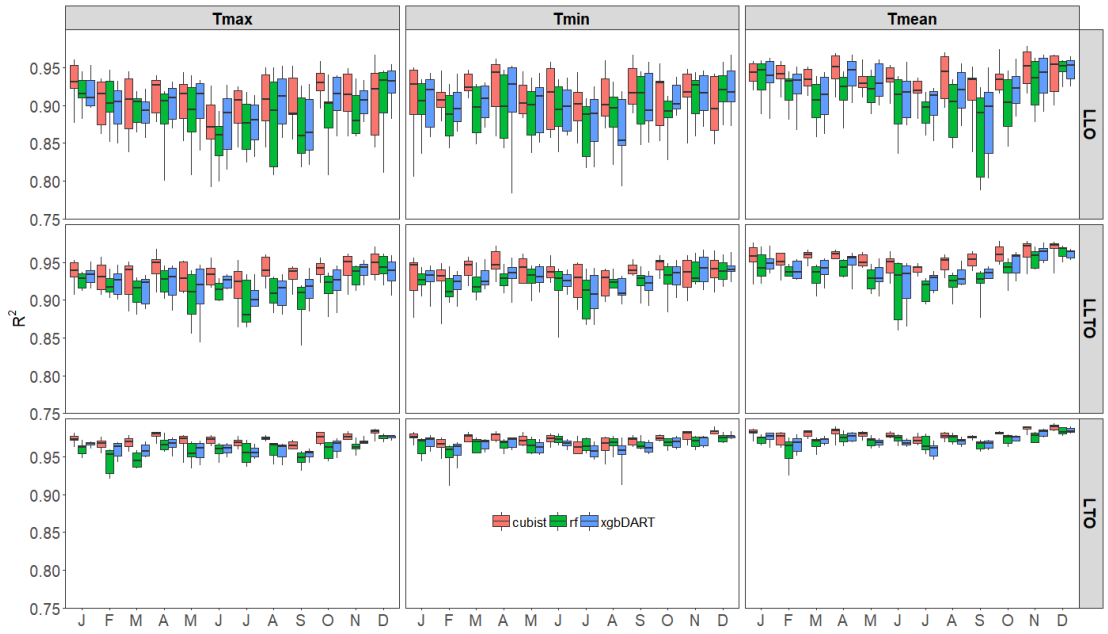


Figure 3-4 LST maximum and minimum temperature validation with in-situ LST measurements in A'rou station

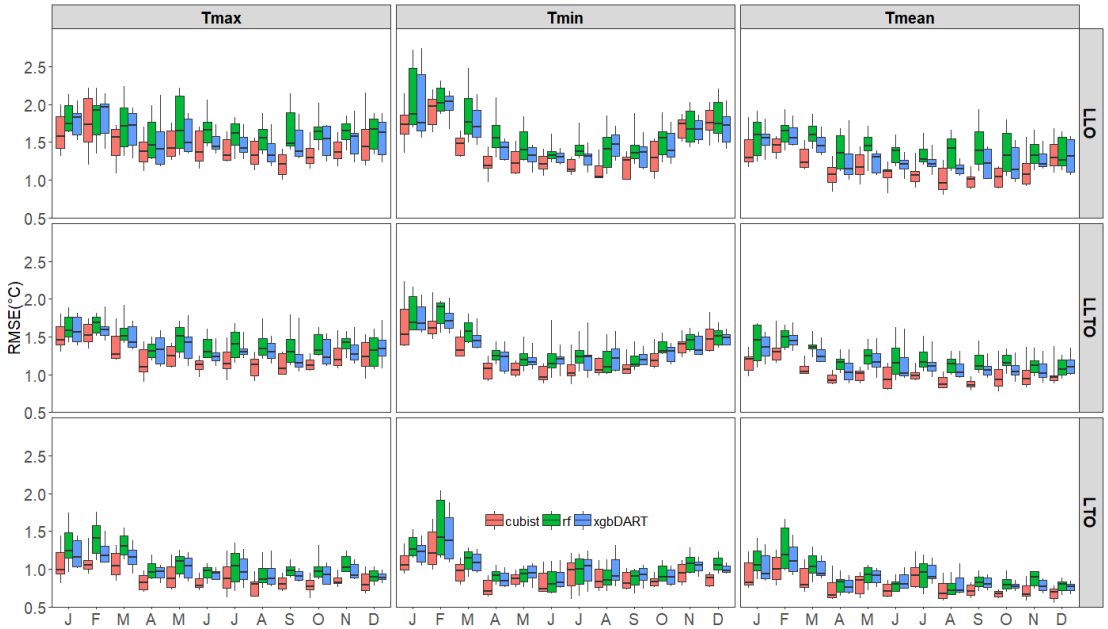
3.3.2 Model performance and variable importance

Figure 3-5 shows that model performances differed from the different temperature products and the target-oriented validation methods. All the 27 models used in this study appeared to have strong relationship ($R^2 > 0.75$ and RMSE < 2.6°C) at the monthly scale. For the three methods used, it can be clearly seen that Cubist regression always showed higher accuracy than XGBoost and RF, but RF was the most robust one with less outliers (not shown on Figure 3-5, R^2 < 1th percentiles and R^2 > 99th percentiles). In terms of the model performance for three temperature products, the Tmean had the highest accuracy and conversely, the Tmax had the lowest. From Tmax to Tmin to Tmean, the model performance

increased, and the differences between RF, Cubist, XGBoost were getting smaller. Additionally, the model performance of LTO-CV was much better than that of LLO-CV and LLTO-CV. The changes of R^2 for different temperature products across 12 months was not obvious but the RMSE showed apparent seasonal variation, higher in winter and lower in summer. Particularly, the RMSE in August (the median RMSE is about 1.0°C) were much lower compared to other months.



(a)



(b)

Figure 3-5 (a) and (b) show the R^2 and RMSE for maximum (Tmax), minimum (Tmin) and mean (Tmean) air temperatures using rf, cubist and xgbDART methods based on LLTO-CV, LTO-CV, LLO-CV. The boundaries of box mark the 25th and 75th percentiles; the horizontal black lines within the box indicate the median; the upper and lower whiskers mark the 90th and 10th percentiles.

Variable importance for each Tmax, Tmin, and Tmean was determined by different Machine Learning methods based on LLTO-CV (Figure 3-6). For all properties in all plots, LSTnight explained most of the variable importance. LSTday was identified as being of major importance to the RF and XGBoost, but LSTday within the Cubist variable importance plots had limited influence on Tmax and Tmean, and its importance even dropped to zero in relation to Tmin. Meanwhile, the importance of elevation demonstrated by Cubist become progressively weaker from Tmax to Tmean, and Tmin. By contrast, RF and XGBoost identified elevation as a decisive factor in for temperature estimates. More specifically, elevation occupied the third or the fourth importance in the plots of Random Forest and always ranked the top three within XGBoost plots. However, for the subplots using XGBoost, there was no clear influence for other variables except LSTday, LSTnight and elevation where their total importance was higher than 95%. Unlike the XGBoost, other indices in the variable ranking plots demonstrated by Random Forest were also good predictors, while both LSTday and LSTnight represents great ability to improve the air temperature estimates. Therefore, these results revealed that RF had greater explanatory capability for temperature estimates than the other two machine learning methods. Modelled solar radiation which was a predictor of Tair in complex terrain also had its explanatory power in the top five within RF. The ranking results of other remotely sensed indices such as EVI and NDSI were expected, because vegetation had a close relationship with temperature and snow cover change reflected by NDSI was an important factor indicative of warming/cooling climate.

Figure 3-6 Tmean, Tmin, and Tmax temperature residuals showed varying temporal sensitivity to physiographic drivers. Each variable was scaled to a total of 100%.

3.3.3 Spatial distribution of surface air temperature

As the monthly time step is more important for many long-term natural resource models, here we focused on the spatio-temporal characteristics of monthly Tmean in 2003-2013 over the TP using both station network and our modelling product. Spatially, the range of Tmean was from -15 ℃ to 20 °C across the TP (Figure 3-7). The warmest month was July and the coldest was January. Monthly Tmax and Tmin (Figure 3-9, Figure 3-10), generally followed the similar spatial patterns, the northwest TP was the coldest area, gradually increased toward the southeast, showing a stepped distribution. As areas in the south and east of the plateau are covered with lush vegetation and strongly influenced by the monsoon, they are amongst the hottest parts of the plateau. Meanwhile, the south eastern part of the plateau was the warmest area followed by the Qaidam Basin in the northeast because of its relatively low altitude. In addition, we derived the seasonal mean temperature and annual mean temperature based on monthly mean temperature over the TP (Figure 3-8), with results similar to that of Zhang et al. (2016).

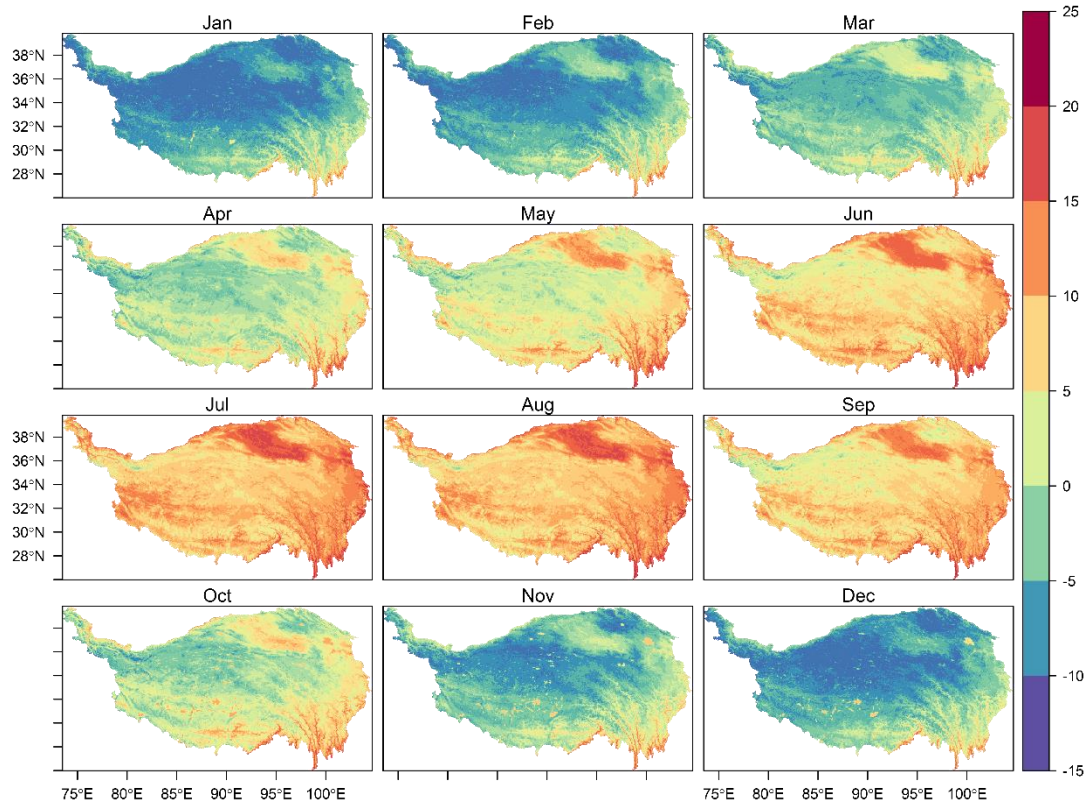


Figure 3-7 Monthly Tmean based on RF model in 2003-2013

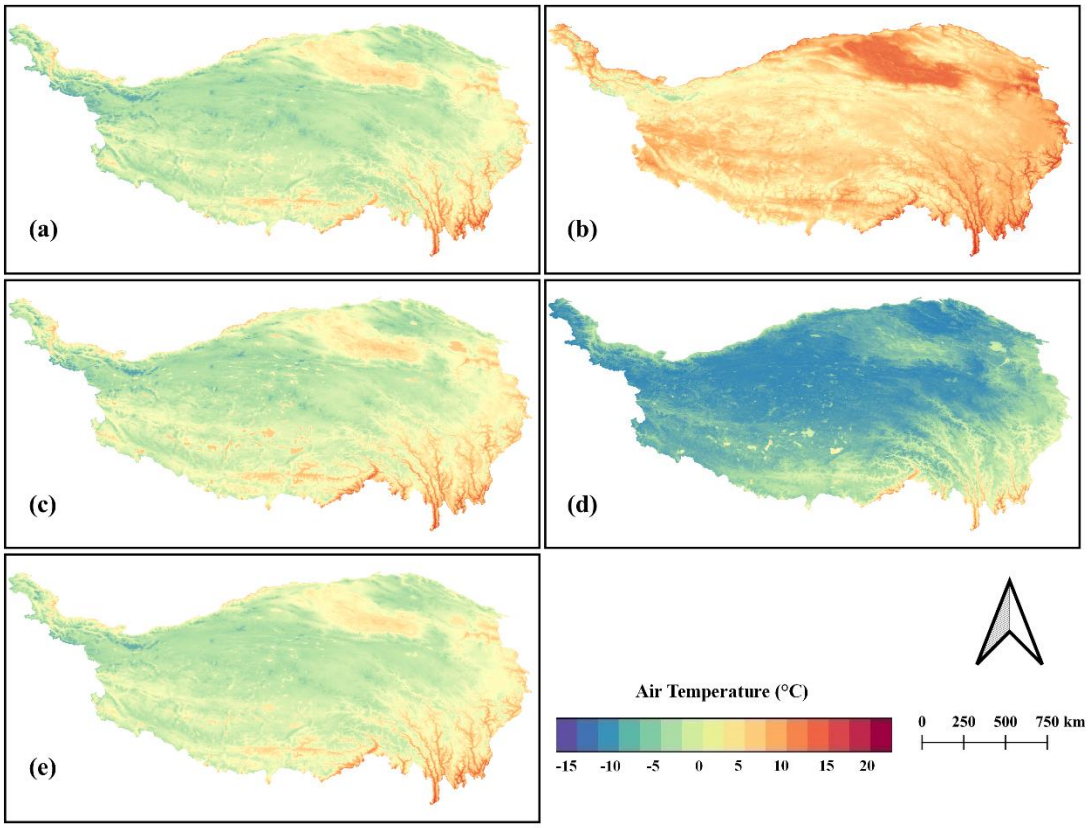


Figure 3-8 Spatial distribution of the seasonally averaged daily mean air temperatures for 2003-2013 in spring (a), summer (b), autumn (c), winter (d) and the full year (e)

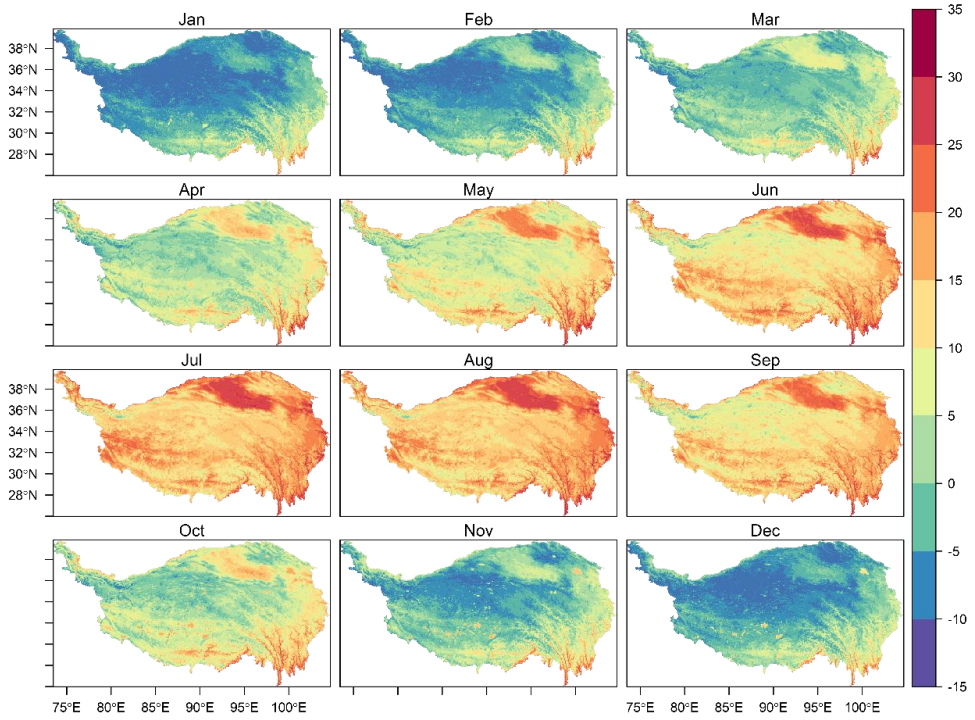


Figure 3-9 Monthly Maximum temperature derived from RF between 2003 and 2013

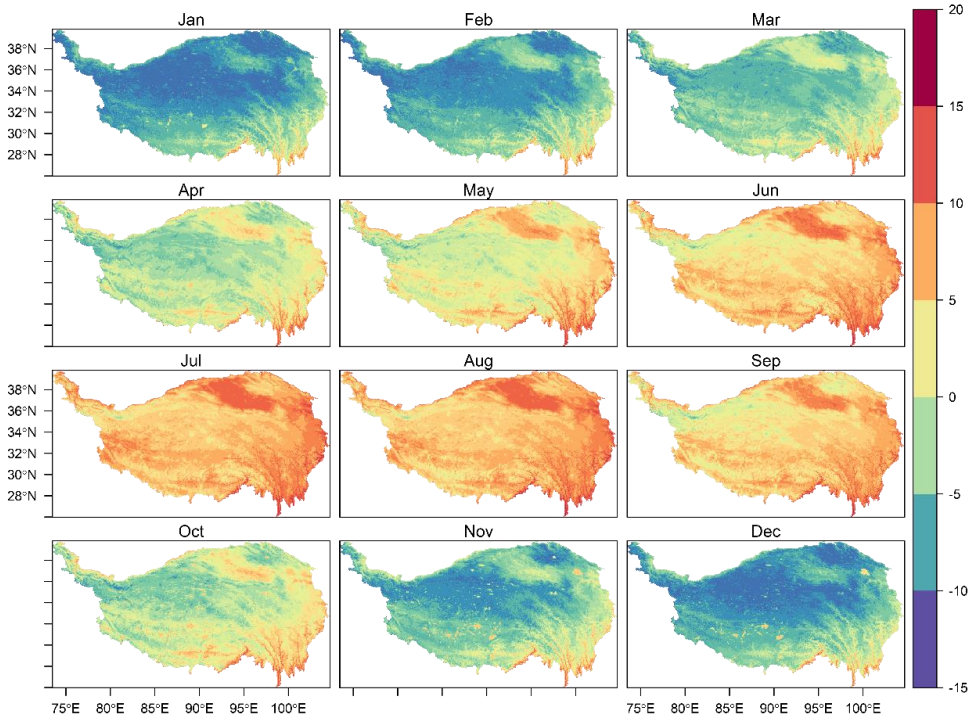


Figure 3-10 Monthly Minimum temperature derived from RF between 2003 and 2013

3.3.4 Comparison with other Tibetan Plateau temperature products

The solar elevation angle in May is relatively high, and daytime ventilation and atmospheric mixing are generally great. By contrast, the lower solar elevation angle, longer night-time and more frequent radiative cooling in December result in colder air drainage and temperature inversions (Daly et al., 2009). Thus, we used Tmean in May to compare the difference between our temperature product and others. Figure 3-11 shows the R^2 using different monthly air temperature products for May and December over the eastern plateau with a dense concentration of climate stations. The Random Forest method resulted in R^2 values of 0.84 and 0.97 for May (Figure 3-11(a)) and December (Figure 3-11(b)) respectively. The monthly Tmean based on TerraClimate products had R^2 values of 0.79 for May and 0.91 for December. In contrast, CMFD showed a lower performance with R^2 of 0.55 for May and 0.78 for December. Thus, the accuracy of our product was better than other temperature products over the eastern TP.

The spatial maps of Tmean for May and December based on different products were shown in Figure 3-12(a). Meanwhile, the density plots from every pixel of those spatial maps were provided for comparison (Figure 3-12(b)). In May, a normal distribution of single peaks appeared; however, in December, a normal distribution of double peaks was observed. The mean values of the density curves in May were about the same, but the variance was different. In comparison, the difference was notable in the case of December, where mean temperatures from TerraClimate were much lower than that from our product. The density of mean temperature in December was in the range of -30 °C to -5 °C. However, according to Zhang et al. (2016), the mean temperature in winter for TP is mostly above -12°C. This difference is partially due to the different models applied. In terms of Tmean generated from Random Forest, it tended to be warmer than the others. For the method applied in TerraClimate, Tair in various regions of the world always has a positive relationship with elevation. According to the research of Cai et al. (2017), Tair at the spring of TP has a negative elevation dependency while summer has a positive elevation dependency. Based on the above conclusion, it is possible to understand the reason that Tair in May shows no obvious differences between the three products.

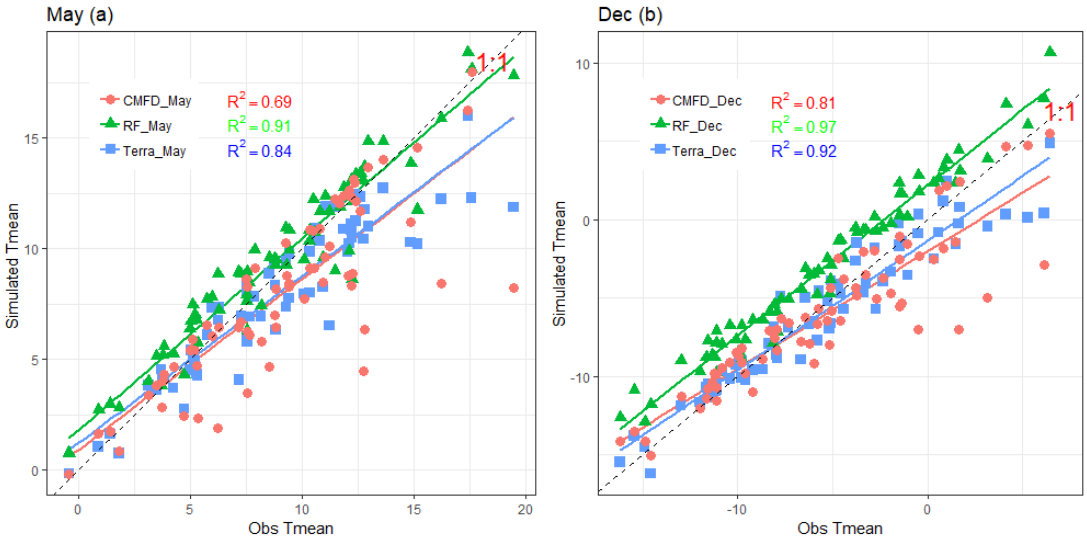


Figure 3-11 The comparison of monthly Tmean in May (a) and December (b) derived from Random Forest, TerraClimate and CMFD with the observed mean temperature of 1980-2010 from in-situ measurements.

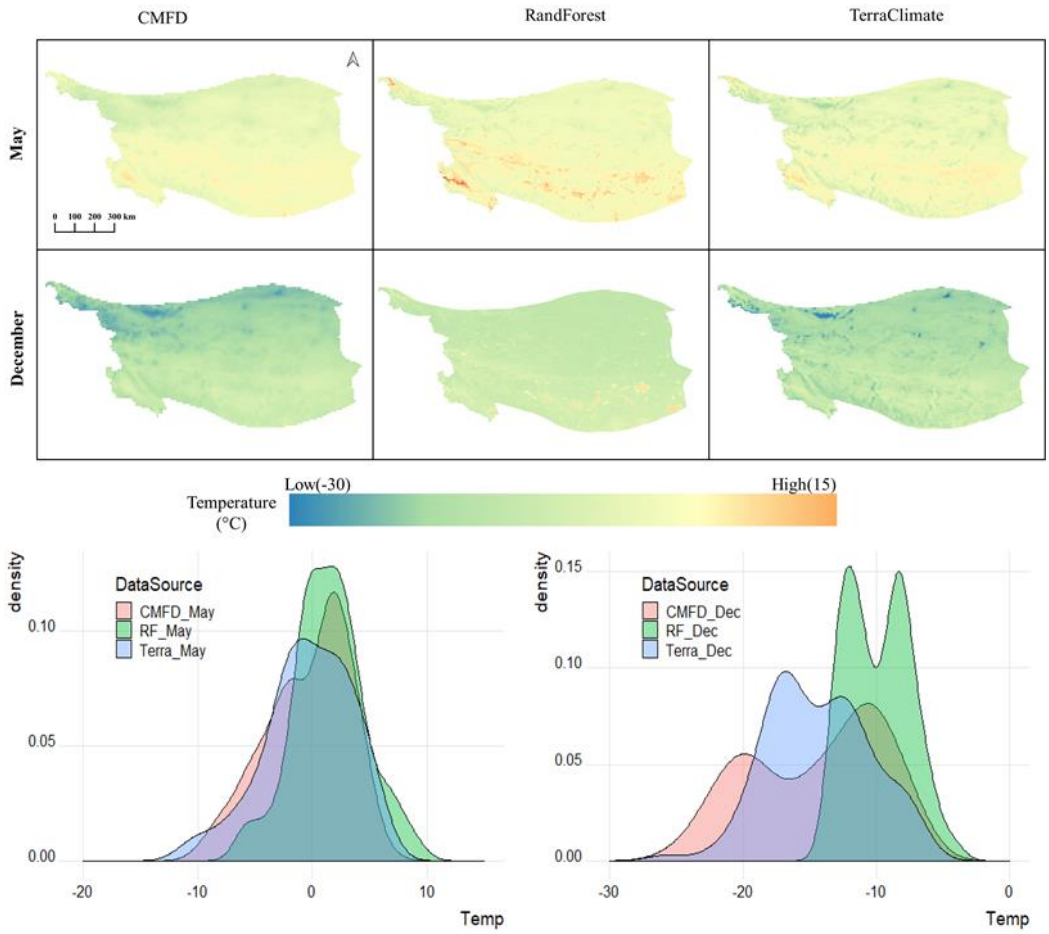


Figure 3-12 (a) Spatial average maps and (b) histograms of Tmean in 2003-2013 at Central TP for May and December

3.3.5 Elevation-dependent warming

Mean Tair products with the highest accuracy is used to explore the evidence of elevation-dependent warming over TP based on three representative elevation zones: 6000-7000 m,4000-5000 m, and 2000-3000 m. As shown on Figure 3-13, in the period of 2003 to 2007, temperature increases in the elevation zones of 2000-3000 m and 4000-5000 m, while temperature decreases at 6000-7000 m. According to the seasonal Mann-Kendall test, only the temperature data in 6000-7000 m (p.value = 0.003) shows a significant cooling trends, while the 2000-3000 m (0.37) and 4000-5000 m (0.11) do not pass the test. In the long period of 2003 to 2013, we found that the cooling trend was observed over 6000 m but the trend at 2000-3000 m and 4000-5000 m are not significant. Furthermore, seasonal Sen's slope test was used to detect the trend, we found

that the magnitude of trend of 6000-7000 m (slope = -0.09) shows negative trend, while both the 2000-3000 m (slope = 0.02) and 4000-5000 m (0.04) show positive trend.

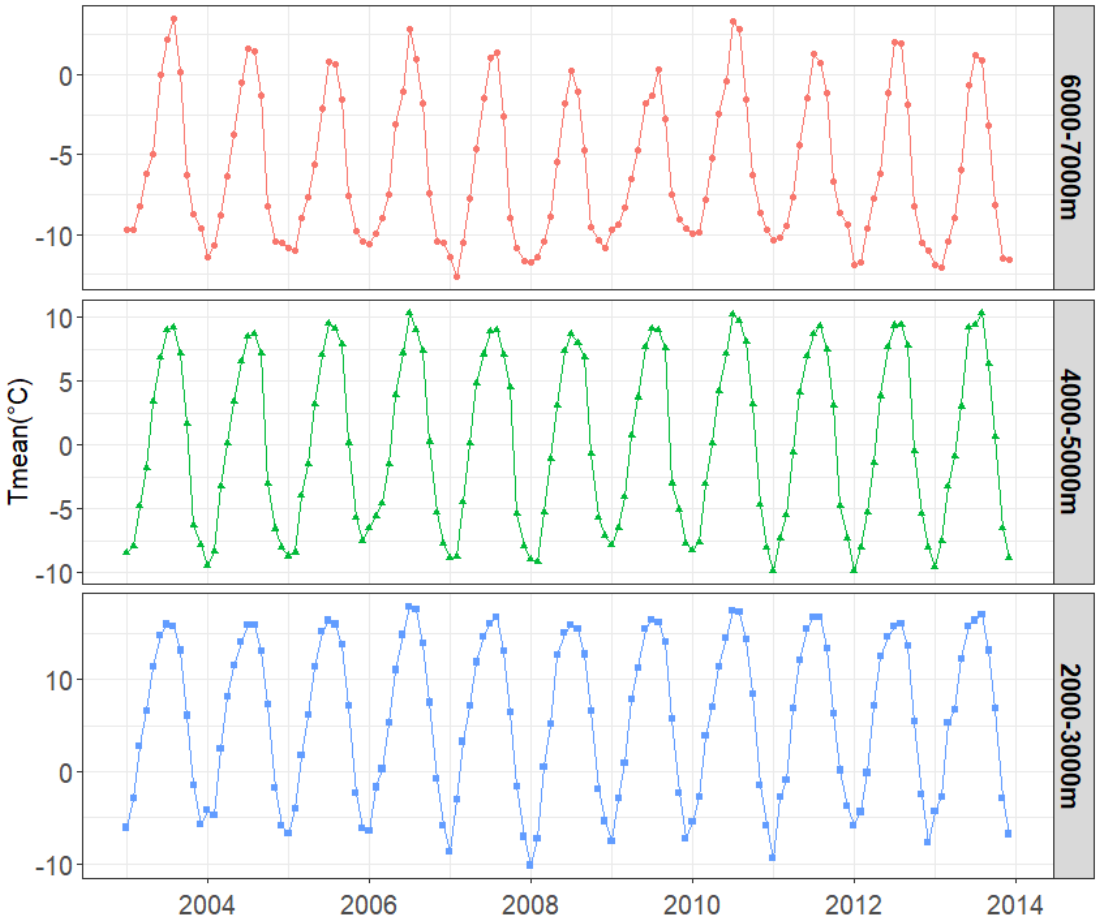


Figure 3-13 Tmean variation at 3 elevation zones from 01/2003 to 12/2013. The number of pixels within 1000 m elevation interval were extracted and each temperature change was computed from the mean of the pixels.

3.4 Discussion

In this study, we investigated the usefulness of a hybrid methodology to provide continuous remote sensed LST for modelling Tair. The application of this methodology over the TP provides thorough datasets in data sparse areas with high elevations by daily MODIS Terra/Aqua LST merging and spatio-temporal gap-filling. It is noteworthy that the missing pixels in the daily LST images after processing by the hybrid method still exist along the south-

eastern boundary of the TP, where the cloud-days are much more frequent than clear-sky days. Key map and figure results were illustrated in a web application (Hybrid MODIS LST Composite Online Tool). This web tool gives us an instant access to global-scale LST data archive and cloud computation capability of GEE. Moreover, with this app link, the user can even acquire the results from mobile devices and does not need to install any software and third-party packages, as opposed to desk applications. Other prospects for the application of this tool include obtaining high-resolution LSTs which would capture the urban heat island phenomenon at a fine scale dynamically. However, Leihy et al. (2018) found great uncertainty in gap-fill predictions for missing LSTs at high elevation sites. For spatially and temporally neighboring predictions, further research needs to consider advanced spatio-temporal gap filling methods that account for using the unique parameters of pixel-specific gap patterns to fill in the missing values (Kong et al., 2019), especially at mountainous area.

The daytime and night-time LST validation results (with an R^2 of 0.75) in this study had a comparable performance with Ouyang et al. (2017), although they used AATSR products at A'rou station for validation. However, due to the difficulty of obtaining representative LST data from ground LST measurements in TP, validation of LST at only one site (A'rou) cannot be deemed representative of the whole TP. Therefore, with the Hybrid MODIS LST online composite tools, more in-situ LST measurements if available can be used for validating in the future. Additionally, the spatio-temporal model validation strategies in this study do not rely on random k-fold cross validation. Instead, stricter validation strategies such as LTO, LLO, and LLTO are used for comparison. However, the aim of this study is to assess the error in both time and space, LLTO CV is finally used for the final target-oriented CV. According to Meyer et al. (2018), Forward Feature Selection (FFS) in conjunction with LLTO allowed removing variables that led to overfitting. For example, the terrain related variables may contribute to overfitting as these "static variables" are overrepresented in the predictor datasets. Concerning the FFS was time consuming for training a large number of datasets and did not show strong evidence to improve model performance and maybe deplete the potential variables to predict

in space and time, FFS was therefore not adopted in this study, but care must be taken when choosing the variables for training. We avoid the overoptimistic view on temperature prediction by using the target-oriented validation (in this case LLTO), the temperature products are still afflicted with errors coming from the characteristics of machine learning algorithms which are not able to predict extreme values (i.e., very low and very high temperatures). Further work is still needed to improve these spatio-temporal models to capture extreme or abnormal temperature in both spatial and temporal domains.

Although in situ observation data below 4000 m clearly show a significant warming trend in the TP, the rate of warming at high-elevation mountains are still unknown. As far as we know, the gridded Tair products in the world are commonly interpolated by spatio-temporal interpolation based on meteorological stations, even in those areas like TP with sparse network of weather stations. In other words, unknown Tair at other locations (such as high mountainous areas and valley areas) is estimated by what we have already known at locations. The key to the success and applicability of common spatio-temporal interpolation are the underlying assumptions employed in describing the relationships and the way in which how these relationships are characterized (Li et al., 2018a). The underlying assumptions of common spatio-temporal interpolation to estimate unknown Tair is that the values of target Tair are spatially autocorrelated. Locations which are closer would have more similarity values of Tair than locations are further apart. Therefore, geographical variables such as elevation and distance are then used to capture and represent these spatio-temporal correlations through geostatistical methods (such as Kriging, IDW). However, geographical variables are consistently static and do not generally capture the temporal dynamics of the spatial pattern of the Tair and cannot represent the effects of biophysical characteristics (Li et al., 2018b). LST was used a proxy for Tair and had been successfully applied to estimate Tair for various regions of the world due to the ability to describe the surface-atmosphere energy exchange process. In this study, we used not just the widely used auxiliary datasets (i.e., LST, locations, elevation, solar radiation and remotely sensed indices) to explicitly represent topographical and

biophysical factors on Tair, but also adopted machine learning algorithms to handle the non-linearity and highly correlated variables. Therefore, the proposed methods are technologically interoperable and scientifically rigor for estimating the Tair at mountainous areas with high elevations. Improved satellite-based temperature Tair products may provide an evidence of elevation-dependent warming. Validation results (Figure 3-11) indicate that the monthly gridded temperature maps show a potential to provide long-term Tair over TP compared to other independent coarse products. Such a product is of great necessity for data scarcity area, which is critical for climate change research. As shown on Figure 3-13, before the year of 2007, temperature increased in the elevation zones of 2000-3000 m and 4000-5000 m and decreased over 6000 m. This finding is consistent with findings in other research (Qin et al., 2009). However, this result is debatable because the period of analysis is extremely short. During the long period of 2003 to 2013, we find that the cooling trend over 6000m was detected by seasonal Mann-Kendall test, and the warming patterns between 2000-3000m and 4000-5000m are not obvious. This phenomenon was also observed from (Pepin et al., 2019). For instance, an increase in temperature was observed in parts of mountainous regions around 4500-5500 m, whereas other mountainous regions observed limited rise of temperature. However, according to previous landscape-scale research on Tair in mountain environments (Todd R and Dean L, 2003), it suggests that significant variation in temperature is occurring at elevation intervals of less than 1000 m. Minimum air temperature, with its strong sensitivity to cold-air drainage, is likely to vary at scales less than 200 m. It is believed that future improvements need to combine the distinct microclimates in high-mountain regions with high-resolution satellite-based datasets.

Unlike plain zones, which are relatively homogenous, mountain areas suffer from local variability, and thus making it extremely difficult to be sure the model simulation are perfect. Notwithstanding these limitations, the factors used for model simulation in this study have been shown to reflect temperature changes well. We find the predictor of NDSI has a higher importance ranking than some other indices. As TP is one of the most sensitive areas to snow

feedback on Earth, the shortened or prolonged snow cover duration in TP will influence the surface air temperature. You et al. (2016) gave a further explanation that the rapid warming in TP is in line with the decreasing snow cover, and hypothesized that change in snow/ice cover exposes the soil to wind or increase the snow line position and alter the absorption of solar radiation, thus leading to the change of surface temperature. Meanwhile, incoming solar radiation also plays a relatively important role at Tair estimation. Additionally, our results (Figure 3-5) show the performance of models estimating Tmin with night-time LST is better than that estimating Tmax using daytime LST. This difference can be partially explained by the fact that night-time LST is more stable than daytime LST. It is also of note that the model simulations are influenced by other climate factors and human activities, such as wind, relative humidity, cloud cover and land use change. However, those factors are not employed due to limited observations. Therefore, further studies by considering more informative indices are needed to improve the model simulation. For example, cloud cover over the TP shows a significant increasing trend; supporting evidence is that there is a significant amount of atmospheric brown cloud generated by fossil fuel consumption and biomass burning over the Indian subcontinent and Asia and be transported to the TP by atmospheric circulation (Ramanathan et al., 2007). As we know, the presence of clouds interferes with the estimation of Tair in mountainous areas. That is probably part of the reason why the accuracy of the estimated Tair from end of summer to end of spring of the following year is not good (Figure 3-5). The most straightforward effect of cloud cover is that MODIS daily LST datasets suffer from a large amount of missing values because of clouds and other atmospheric conditions (see in Figure 3-3(a), Figure 3-3(c)). Therefore, we adopted the hybrid approach (combine several methods) to fill the gaps in high spatio-temporal LST datasets before using them for estimating Tair. However, due to the lack of the cloud cover information, there is still a deviation between the gap-filled LST and the real LST on cloudy days. Another indirect effect of cloud cover is the estimation of other predictors (i.e., incoming solar radiation in TP). Instead of using cloud-based model, a sunshine-based model was adopted to simulate incoming solar radiation under the influence of clouds. Therefore, if appropriate atmospheric parameters for the study area,

such as cloud cover, atmospheric transmissivity and atmospheric turbidity can be obtained, the final modelling results would be better. In addition, factors like wind and relative humidity data also contribute to the estimation of Tair. For example, weakening of zonal wind speed also raises temperature in the Qaidam Basin significantly (Wang et al., 2014b).

3.5 Conclusions

In this study we build an online tool based on a MODIS LST "hybrid" methodology to generate continuous daily maximum and minimum land surface temperature datasets in locations without observations and to provide the required remotely sensed inputs to air temperature prediction models. Changes in received solar energy among mountains inevitably affect the earth's energy budget. We integrate mountain solar radiation and diverse remotely sensed vegetation indices to provide reliable temperature products over the TP. By comparing the performance of different machine learning techniques, we found the RF model performed best in predicting Tmax, Tmin, and Tmean. We expect the methodology we have developed can be potentially useful for improving temperature datasets in mountainous regions around the globe, and thereby also improving climatic, environmental, hydrological and ecological models.

Chapter 4. Heat wave tracker: a multi-method, multi-source heat wave measurement toolkit based on Google Earth Engine

This chapter is based on the following manuscript:

Zhang, M., Yang, X., Cleverly, J., Huete, A., Zhang, H., & Yu, Q. (2021). Heat wave tracker: A multi-method, multi-source heat wave measurement toolkit based on Google Earth Engine. *Environmental Modelling & Software*, 105255.

Abstract

Under ongoing global warming due to climate change, heat waves in Australia are expected to become more frequent and severe. Extreme heat waves have devastating impacts on both terrestrial and marine ecosystems. A multi-characteristic heat wave framework is used to estimate historical and future projected heat waves across Australia. A Google Earth Engine-based toolkit named *heat wave tracker* (HWT) is developed, which can be used for dynamic visualization, extraction, and processing of complex heat wave events. The toolkit exploits the public long-term high-resolution climate datasets to developed nine heat wave datasets across Australia for extreme heat wave value analysis. To examine climate change on heat waves and how they vary in time and space, we also explore the probability and return periods of extreme heat waves over a period of 100 years. The datasets, toolkit and findings we developed contribute to global studies on heat waves under accelerated global warming.

Key words: *Extreme heat wave; Google Earth Engine; climate datasets; risk analysis; GCM; Australia*

4.1 Introduction

Under ongoing global warming due to climate change, heat waves are expected to become more frequent and severe in the future (IPCC, 2019). Extreme heat waves during the last two decades have been recorded across many regions in the world such as those in Europe in 2003 (Schär et al., 2004), Moscow region in Russia in 2010 (Rahmstorf and Coumou, 2011), and Australia in 2013 (Lewis and Karoly, 2013). Heat waves in Australia incur significant hazard for both humans and ecosystems and cause more deaths than other natural hazards including floods, storms and bushfires. In terms of heat wave impacts on ecosystems, extreme heat waves increase the probability of bushfire risk, affect crops and food security for terrestrial systems (Luo, 2011), and cause catastrophic damage to marine ecosystems (Hobday et al., 2016). Moreover, extreme temperatures contribute to widespread unfavorable health outcomes and even the death of vulnerable people.

Although heat wave is commonly known as a period of exceptional hot weather event, there is currently no universal informative measurement in climate science community (Alexander and Perkins, 2013). To overcome these issues, a set of climate indices developed by the Expert Team on Climate Change Detection and Indices (ETCCDI) have been widely applied to observational and modelled climate data to understand previous and future changes in extreme heat wave events (Alexander et al., 2006; Zhang and Yang, 2004). The work by ETCCDI is extensively recognized as pioneering, however, the indices only measure one feature of extreme events such as frequency, intensity or duration (Perkins, 2015). A comprehensive and consistent analysis of heat waves is required, which should consider multi-characteristics of heat wave events, namely: i) frequency, ii) intensity, iii) duration, and iv) spatial extent (Raei et al., 2018). The multi-characteristic heat wave definition method used in this study is from a well-known heat wave framework constructed by Alexander and Perkins (2013) and includes: a minimum temperature approach, a maximum temperature approach, and an excess heat factor (EHF) approach. This framework has proven to be successful in measuring historical and future projected heat waves.

However, useful public software or tools that identify all the required characteristics of heat waves (frequency-intensity-duration-spatial extent) are still rare. Most studies with their own tools cannot fully reflect the four characteristics of complex heat wave events (Feron et al., 2019; Li, 2020; Lyon et al., 2019). By summarizing the classical heat wave definition, an R package called heatwaveR was developed, which provides a comprehensive analysis to detect and visualize ocean heat waves (Schlegel and Smit, 2017). However, it is inefficient when applied to large gridded data products. Global Heatwave and Warm-spell Data Record and Analysis Toolbox (GHWR) which is a MATLAB Toolbox allows processing and extracting heat wave records for any location efficiently. It not only contains multiple definitions but also detects the required multi-characteristics (Raei et al., 2018). However, desktop applications like GHWR still have a bottleneck when encountering the challenges related to accessibility of long-term gridded climate data, data storage and computational requirements. In the current era of big spatial and Earth Observation (EO) data, users need to deal with a vast amount of different spectral, temporal and spatial resolutions data (Gomes et al., 2020). To meet these demands, there is need for novel technologies based on cloud computing to properly extract heat wave information in the server side without having to download vast amounts of climate data and provide dynamic visualization, extraction and processing of complex heat wave events. Google Earth Engine (GEE), a powerful cloud computing geospatial analysis platform, has given researchers the opportunity to use big data for petabyte-scale environmental data analysis (Gorelick et al., 2017a).

With the gridded global reanalysed datasets (e.g., Hadley Centre/Global Historical Climatology Network (HadGHCND), Climate Prediction Centre (CPC)) and regional reanalysis datasets (e.g., The COordinated Regional Downscaling EXperiment (CORDEX), Australian Water Availability Project (AWAP)) being freely available, many studies have investigated heat waves at various scales (Christidis et al., 2014; Ma et al., 2020; Perkins et al., 2012). The atmospheric reanalysis datasets are quite useful for gaining understanding in how the heat wave will change. Reanalysis datasets are created by data assimilation and numeric

models to represent a synthesized estimate of the atmospheric state and provide global scale dataset over several decades or longer. One benefit of using reanalysis data is that it extends the study to locations without observation records. Another important advantage is that the spatially contiguous heat wave regions derived from the reanalysis data have crucial implications for heat-related impacts such as exposure of the community to extreme heat wave events and high energy demands (Li, 2020; Lyon et al., 2019). However, some heat wave assessments are mostly based on climate datasets with relatively coarse resolution which would affect the representation of heat waves, resulting in biased conclusions. Furthermore, key processes that occur on regional scales may not be adequately simulated. Benefiting from those newly reanalysed climate datasets and high spatio-temporal gridded regional climate datasets, our analysis will explore how these climate datasets differ in representing heat waves and how the methods differ in identifying and characterizing heat waves.

Increasingly, researchers are becoming less interested in data in the "normal" range and more concerned with the 'abnormal' and extreme events that are recurrent and unpredictable. Extreme value theory (EVT) is the statistical framework that estimates the probability of an extreme event occurring in the future (Coles et al., 2001). Because of its importance, many public packages and toolboxes over the last decade have been developed to implement various methods from EVT (Cheng et al., 2014; Gilleland and Katz, 2016; Heffernan et al., 2016; Ribatet et al., 2011). It is clear from much of the literature using gridded observed data and projected climate model data at regional and global scales that the probability of extreme heat waves will change over time (Alexander and Perkins, 2013; Purich et al., 2014). Recently, several studies of the risks of heat wave by means of the EVT have been published (Ma et al., 2020; Shen et al., 2016; Tanarhte et al., 2015). However, the precise probabilities of intensity, frequency and duration of extreme heat wave at a continent scale like Australia over the time are still unknown. Meanwhile, the potential impact of climate change on heat wave varies in space and time. In this context, we can explore the risk of heat waves in Australia by performing non-stationary analysis of extreme heat waves for the past 100 years.

In this study, we will develop a multi-method global heat wave data record and analysis toolbox (namely heat wave tracker) to process and extract heat wave records from multi-source climate datasets. The core algorithms behind the toolbox are based on a general heat wave framework which employs three separate heat wave identification methods (daily minimum and maximum temperature, and the excess heat factor) and use a fixed threshold as the baseline to determine a heat wave event which has at least three days in a row where the threshold is exceeded. With our toolbox's computational power in handling long-term high-resolution climate datasets, we developed nine extreme heat wave datasets in Australia for extreme heat wave value analysis. In addition, we first use non-stationary generalized extreme values method to analysis the characteristics of extreme heat wave events in Australia over the past 100 years to help adjust policies for climate change adaptation. Finally, we also explore how the characteristics of heat waves are projected to change across Australia under future climates.

4.2 Data and methods

4.2.1 Earth observation datasets

SILO is a database of Australian climate data from 1889 to the present hosted by the Queensland Department of Environment and Science (DES) (https://www.longpaddock.qld.gov.au/silo/). It provides daily climate variables on a 0.05° grid across Australia for research, modelling and climate applications. The datasets are constructed from observational data obtained from the Australian Bureau of Meteorology (BoM). SILO uses a thin plate smoothing spline to interpolate daily climate variables. There is some evidence that the data quality of maximum and minimum temperatures corresponds strongly to station density, with the largest errors tending to occur where the network of observed stations is sparse (Jones et al., 2009). Currently, SILO data are uploaded into the GEE data catalog and maintained by Earth Observation Data Science (https://www.eodatascience.com/).

In addition to using high-resolution interpolated climate data, there have been many studies

using reanalysed temperature data for heat wave studies, such as the latest fifth generation ECMWF (European Centre for Medium-Range Weather Forecasts) reanalysed climate data (ERA5) and CPC Global Daily Temperature dataset dating back to 1979 (https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html). ERA5 combines physical modelling and data assimilation into a complete hourly-based and consistent dataset. For example, minimum and maximum daily air temperature at 2m from EAR5 Daily are calculated based on the hourly 2m air temperature data. The ERA5 Daily used in this study were obtained within the GEE Data Catalog (https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_DAILY). CPC Global Daily Temperature dataset includes both daily Tmax and Tmin on a 0.5*0.5 grid from 1979 to the present. This product is constructed by a combination of two weather station datasets around the world, namely Climate Anomaly Monitoring System (CAMS) and Global Historical Climatology Network version 2 (GCHN). These two datasets together have about 10978 stations around the global, the temperatures from which are gridded using Inverse Distance Weighting (IDW) interpolation algorithm. In addition, the temperature lapse rate estimated from observation-based global reanalysis temperatures are used to make topographical adjustments. Note that observations from CAMS and GCHN have less coverage over central Australia and good coverage over USA, Europe, and China. The lack of accuracy from the sparse density of observation stations would impact the identification of heat wave events. In this study, CPC dataset netCDF4 files have been transformed into GeoTIFFs format using R scripts and uploaded into the GEE Catalogue for further analysis.

For projection periods (2006-2100), Coupled Model Intercomparison Project Phase 5 (CMIP5) that have daily maximum and minimum temperature from the historical experiment and two representative concentration pathway (RCP) experiments (RCP4.5 and RCP8.5) are analysed in this study. Within the GEE data catalog, the NASA NEX dataset contains daily downscaled projections of 21 GCMs under the CMIP5 across two greenhouse gas emissions scenarios (https://developers.google.com/earth-engine/datasets/catalog/NASA_NEX-GDDP).

CMIP5 reference periods (1975-2005) and projection periods (2006-2100) which contain daily maximum and minimum temperature are used to construct multi-model mean composites for summer heat wave under two RCP emission scenarios.

Table 4-1 Datasets used in this study

Dataset Name	Spatial Resolution	Time Period	Data Source
SILO	0.05*0.05	1920-2020, daily	EO Data Science (GEE)
ERA5	0.25*0.25	1979-2020, daily	ECMWF reanalysis climate data (GEE)
CPC	0.5*0.5	1979-2020, daily	CPC global temperature (NOAA)
CMIP5	0.25*0.25	1950-2099, daily	NASA NEX-GDDP (GEE)

4.2.2 Heat wave indices

The core algorithms behind the toolbox are based on a general heat wave framework which employs three separate heat wave identification methods (daily minimum and maximum temperature, and the excess heat factor) and use the fixed and dynamic thresholds as the baseline to determine a heat wave event which has at least three days in a row where the threshold is exceeded. From a climatological perspective, heat wave indices with absolute thresholds such as ETCCDI may only be suitable when studying heat waves in a small region where a single climate regime exists. However, for large regional or continental studies like Australia where a broad range of climates exist, three separate heat wave identification methods used in this study can be readily calculated from climatological data is more applicable for representing heat wave occurrence across multiple climates. Of which, EHF is not only more sensitive than other heat wave indices in measuring heat waves, but is also the official definition used Australia-wide (Alexander and Perkins, 2013; Nairn and Fawcett, 2015). For each grid

point, three heat wave indices were calculated for the Australian warm season from November 2018 to March 2019. These indices include:

1) TX90pct—The 90th percentile of Tmax in calendar day based on a centered 15-day window (i.e., 7 days after and before a calendar day). The thresholds are calculated for each time period and grid point separately. The unit of TX90pct is °C.

2) TN90pct—The 90th percentile of Tmin in calendar day, same time period and unit as Tmax.

3) Excess heat factor (EHF) – EHF is a product of two metrics based on Tmean: EHI_{sig} and EHI_{accl}; The first index is denoted as 'significance' (EHI_{sig}) and determines how extreme the temperature conditions are by comparing the previous 3-day mean with climatology (the 95th percentile of the daily mean temperature calculated over the period of reference) (Equation (1)); The second index is a measure of acclimatization (EHI_{accl}) and the difference of the 3-day mean to the previous 30-day mean (Equation (2)). With this second index, heat stress is only likely to occur in summer. From Fig. 1, the threshold 0 means the unusual 3-day mean temperature is above the 95th percentile of the average temperature over a fixed climatological period. EHF can also be defined as EHF = | EHIaccl | × EHIsig, which means EHIaccl acts as an amplification term on EHIsig, thus EHF can be negative.

$$EHI_{sig} = [(T_i + T_{i-1} + T_{i-2})/3] - T_{95} \quad (1)$$

$$EHI_{accl} = [(T_i + T_{i-1} + T_{i-2})/3] - [(T_{i-3} + \cdots + T_{i-32})/30] \quad (2)$$

$$EHF = EHI_{sig} * \max [1, EHI_{accl}] \quad (3)$$

For heat wave identification method based on daily mean temperature, heat wave represented as excess heat factor (EHF) is a product of two metrics: EHIsig and EHIaccl. So, the unit of heat wave is given in °C2. However, for heat wave identification method based on daily minimum and maximum temperature, heat wave is defined as a spell of at least three consecutive days with daily minimum and maximum temperature exceeding the local 90th percentile of a centered 15-day of window. Therefore, the unit of heat wave is given in °C.

Further to these three indices, we used a multi-aspect framework to represent heat wave characteristics including:

(1) Heat Wave Number (HWN) - the total number of discrete heat wave events;

(2) Heat Wave Duration (HWD) - the length of the longest heat wave event;

(3) Heat Wave Frequency (HWF) - the sum of days satisfying positive EHF;

(4) Heat Wave Amplitude (HWA) – the peak magnitudes (the highest value of the heat wave in a season);

(5) Heat Wave Magnitude (HWM) – the mean magnitudes (average magnitude across all heat waves);

Among them, HWM and HWA are measures of heat wave intensity, while HWD, HWF and HWN are measures of heat wave longevity.

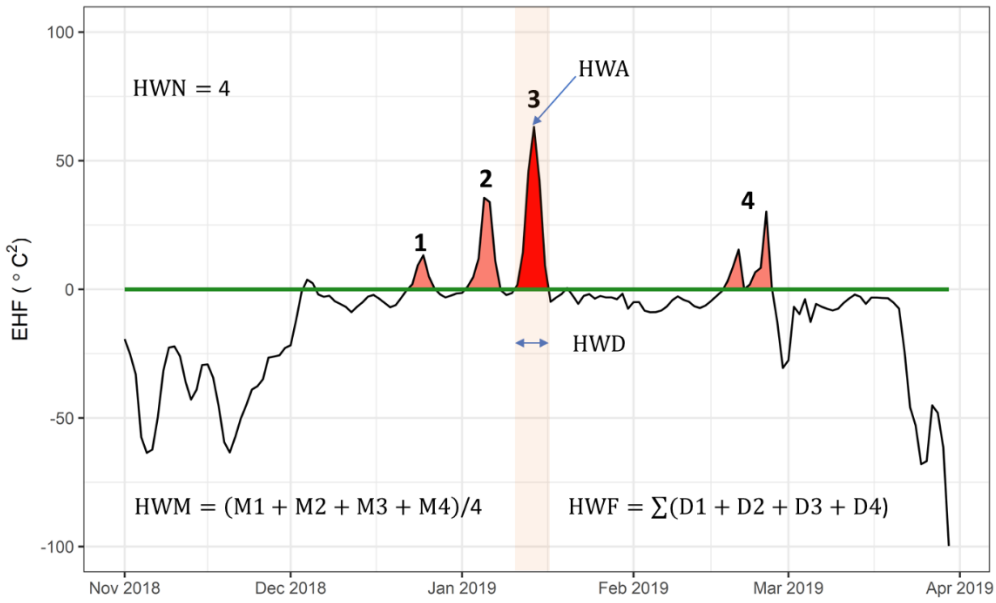


Figure 4-1 An example schematic of indices used to define heat wave-EHF. Short duration heat spikes less than three days in a row are not heat waves. In this figure the green line is the threshold and black line is the EHF. There are four discrete events including red and pink heat spikes (HWN); the highest red heat spikes is the heat wave amplitude (HWA); the length of the longest event is also the red heat spikes (HWD); the average heat wave magnitude is the average

magnitude across four events (HWM); and the sum of four heat wave events that above the threshold is HWF. The five indices in the figure are calculated for each season and annually.

4.2.3 Non-stationary generalized extreme value analysis

Extreme value theory (EVT) has a rigorous framework for analysis of climate extremes and their return levels (Coles et al., 2001). Generalized extreme value (GEV) distribution is a combination of three limiting distributions: Gumbel, Fréchet, or Weibull comes from the limit theorems for block maxima/minima or annual maxima/minima (Katz, 2010). A variety of studies apply the GEV to analyze climatic extremes. This technique is often referred to as the block maxima approach. Another form of the EVT is known as the peak-over-threshold (POT) approach, in which extreme values above a high threshold are analyzed using a generalized Pareto distribution. Both block maxima approach and POT are widely applied in studying climatic extreme events. The cumulative distribution function of the GEV can be expressed as:

$$\psi(x) = \left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{\frac{-1}{\xi}}\right\},\ \left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right) > 0 \quad (4)$$

The GEV distribution has three distribution parameters $\theta = (\mu, \sigma, \xi)$: (1) the location parameter (μ) determines the center of the distribution; (2) the scale parameter (σ) specifies deviations around μ; and (3) the shape parameter (ξ) governs the tail behavior of the GEV distribution. For $\xi > 0, \xi \to 0$, and $\xi < 0$ leads to Frechet distributions, Gumbel distribution and Weibull distribution, respectively.

The extreme value theory for stationary random sequences has been extensively studied. In this study, a stationarity process assumes no change to extreme's properties while a non-stationary process is time-dependent, and the properties of the distribution would change in the future. The location parameter is assumed to be a linear function of time to account for non-stationarity, while keeping the other two parameters constant:

$$\mu(t) = \mu_1 t + \mu_0 \quad (5)$$

where t is the time (in years), and $\beta = (\mu_1, \mu_0, \sigma, \xi)$ are the parameters. In this study, a

practical package named *Non-stationary Extreme Value Analysis* (*NEVA*) Matlab package was introduced for assessing extremes in a changing climate. *NEVA* offers a framework for performing non-stationary analysis of extremes and provides non-stationary effective return levels with t-year return period, and risks of climatic extremes using Bayesian inference and also includes simulated ensembles with upper bound and lower bound (Cheng et al., 2014). This study estimated extremes heat wave metrics based on non-stationary Maximum Likelihood Estimators. Here, from the long term (1920-2019) time series of heat wave magnitudes, non-stationary GEV was fitted together with the standard error using R package *Introduction to Statistical Modeling of Extreme Values* (*ismev)*. We kept the scale and shape parameters constant, while the location parameters were calculated from the regression parameters (μ_1, μ_0) of Equation (5) at the median of the corresponding time period. For example, the median of the corresponding time was 1970 over the period 1920-2019. For the sub-time periods (1980-2019), the estimation for the non-stationary GEV distribution is similar.

4.2.4 Online heat wave measurement under a framework

The heat wave tracker is to facilitate the exploitation of the up-to-date climate data described in Table 1 by providing users a multi-characteristic and multi-source heat wave measurement toolkit. The entire process of heat wave measurement at a continental scale is shown in Figure 4-2. The required inputs for our online system include the historical climate data and their future projection. With long time series of climate data, two separate methods were used to calculate fixed and dynamic thresholds. The fixed thresholds are calculated by the 95th percentile of a fix reference period. The dynamic thresholds are based on the 90th percentile of a temporal moving window. Three separate heat wave indices were then used to determine the heat wave characteristics. The core algorithm contains five iterations, three band math operations and two spatial operations to retrieve five heat wave characteristics at each grid. The first iteration was to do an accumulation of the number of positive values of heat wave indices. The second and third iteration were combined to detect heat wave events, defined as a spell of at least three consecutive days with values of heat wave indices exceeding the threshold. The fourth iteration

was used to find the end point of each heat wave events. The fifth iteration was used to accumulate the positive values of heat wave indices. Based on those extreme value analyses and heat wave characteristics database, we created an online heat wave tracker app for public users.

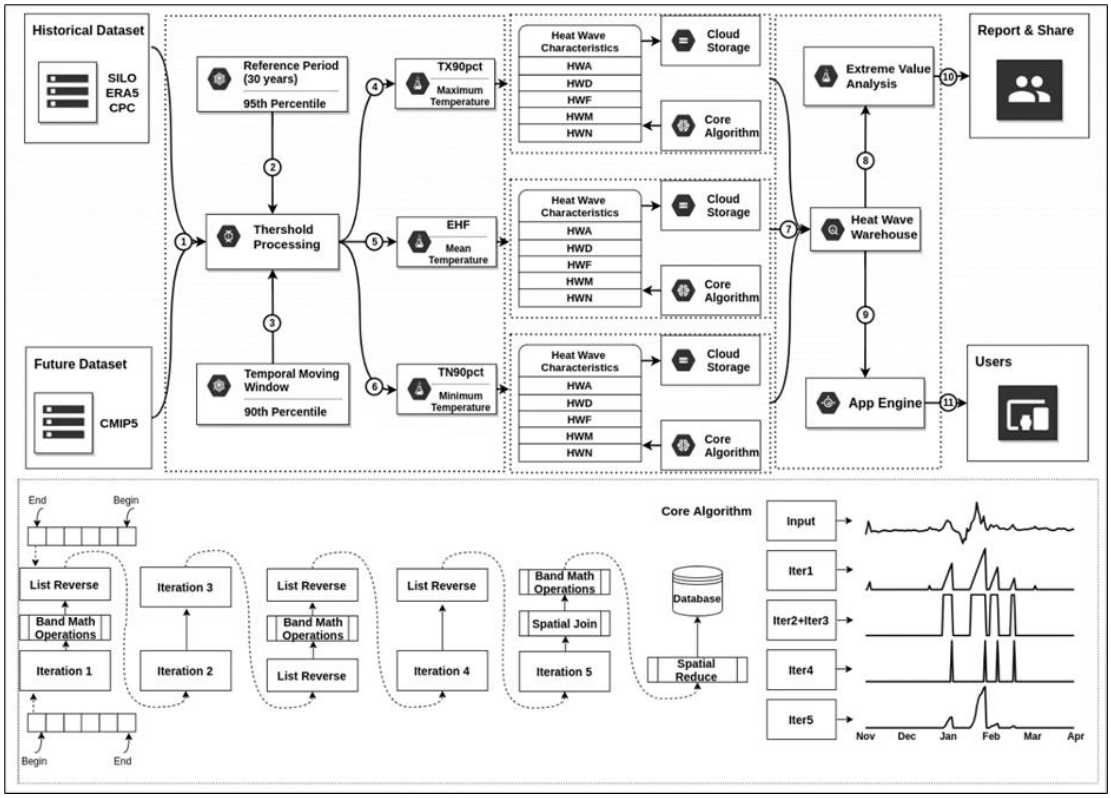


Figure 4-2 The online implementation of heat wave tracker toolkit based on Google Earth Engine, using a framework enables climate data integration for heat wave measurement at a continental scale.

4.3 Results

4.3.1 Heat Wave Tracker

Heat wave tracker is a user-friendly web tool we developed in Google Earth Engine (GEE). The temperature datasets and heat wave definition outlined above are integrated into this online software tool to study heat waves in Australia. The first step is to pre-define the temperature above a certain threshold and pre-process the corresponding five-month long heat wave records.

More precisely, thresholds from the reference period of SILO data (1960-1990) and the reference of ERA5 data (1979-1999) were calculated beforehand. Then, the multi-source heat wave record datasets (e.g., heat wave records between 1990-2019 are from SILO, 2000-2019 are from ERA5, 2000-2019 are from CPC, 2030-2099 are from CIMP5) using multi-method are generated and stored in GEE cloud data catalog for further visualization analysis to decrease processing times. Subsequent steps are performed in the graphical user interface (GUI), the users can define the point of interest and select the year, data type, heat wave identification method and run the program. Then the tool will plot several figures displaying the time-series of heat wave records and five heat wave metrics maps (HWN, HWD, HWF, HWM, HWA). The information can also be exported (e.g., CSV files) for further analysis. In such a case, analysis-ready heat wave records prove to be a practical and economical way for real-time and human-interactive visualization. Heat wave tracker is freely available from the authors for educational and academic purposes at https://github.com/geogismx/Heatwavetracker. The online tool is publicly available at https://tensorflow.users.earthengine.app/view/heat-wave-tracker. While we have focused on the heat waves of Australia, users can also define their own research area and produce their heat wave outcomes. For example, users can even use the tool to evaluate the global heat wave with ERA5 datasets.

4.3.2 How do the datasets differ in representing heat waves?

Despite the use of the same heat wave definition (EHF), different temperature datasets may provide different heat wave metric maps. It relates to the issues of spatial resolution, instrumentation and data quality. An example of the spatial variation from different climate datasets for heat wave metrics identification is given in Figure 4-2, which shows the heat waves across Australia in 2018-2019 (over the period of November-December-January-February-March) from SILO gridded datasets, ERA5 reanalysis datasets and CPC Australia daily temperature datasets. Generally, climate datasets with a high spatial resolution are much smoother than those with lower spatial resolution.

Each heat wave metric between the three datasets shows similar data range on the color

scale. However, the contiguous spatial distribution clearly differs between the three datasets. Specifically, the extreme HWA for each dataset all occur over southern Australia while northern Australia does not experience extreme heat waves. HWA can increase up to $80\,°C^2$ in the northwest of NSW. In ERA5, larger HWA values are more confined to lower elevations of southern Australia, whilst HWA in SILO and CPC also appear in the central areas. Similar to HWA, the spatial pattern of HWM is mainly centered around the south coast and northwest of NSW. However, the anomalous red spots of HWM in CPC may be caused by the coarse resolution. It is interesting to note that the HWF and the HWN are similar but do not always overlap. From these three datasets, we can see that the HWF and HWN are located in north western and southeast Australia. We also find that HWN from CPC can reach up to 12 times per year and is about two times larger than that from SILO and ERA5, implying that caution should exert when using CPC. The HWF has some influences on HWD, which means the extent of HWD almost falls in the regions of HWF.

Since local scale differences cannot be detected by simple visualization or in cell by cell comparison, we used a map comparison method named the structural similarity index (SSI) to identify local differences in terms of mean, variance and covariance between two maps (Islam et al., 2020; Jones et al., 2016; Wiederholt et al., 2019). Based on the global average value of the SSI metric, we try to provide a quantitative analysis of which climate data set is more reliable with respect to five aspects—HWA, HWD, HWF, HWM, HWN. From Table 2, we can see that the similarities between three gridded datasets in terms of five aspects are quite different. There is strong similarity between ERA5 and SILO (0.78) in HWA. The SSI between CPC and ERA5 in HWA is similar (0.68) but weaker for SILO (0.67). The strong level of SSI between ERA5 and SILO (0.77) is also found in HWD, while ERA5 and CPC has a similarity of 0.67, the weakest similarity of 0.66 is from SILO and CPC. The occurrence-based aspects like HWF and HWD lead to reduced similarity. The weaker similarity in HWN exists between three climate datasets, but the SSI between CPC and ERA5 is better (0.56) than CPC and SILO (0.55). Overall, it suggests that ERA5 is the most reliable climate dataset.

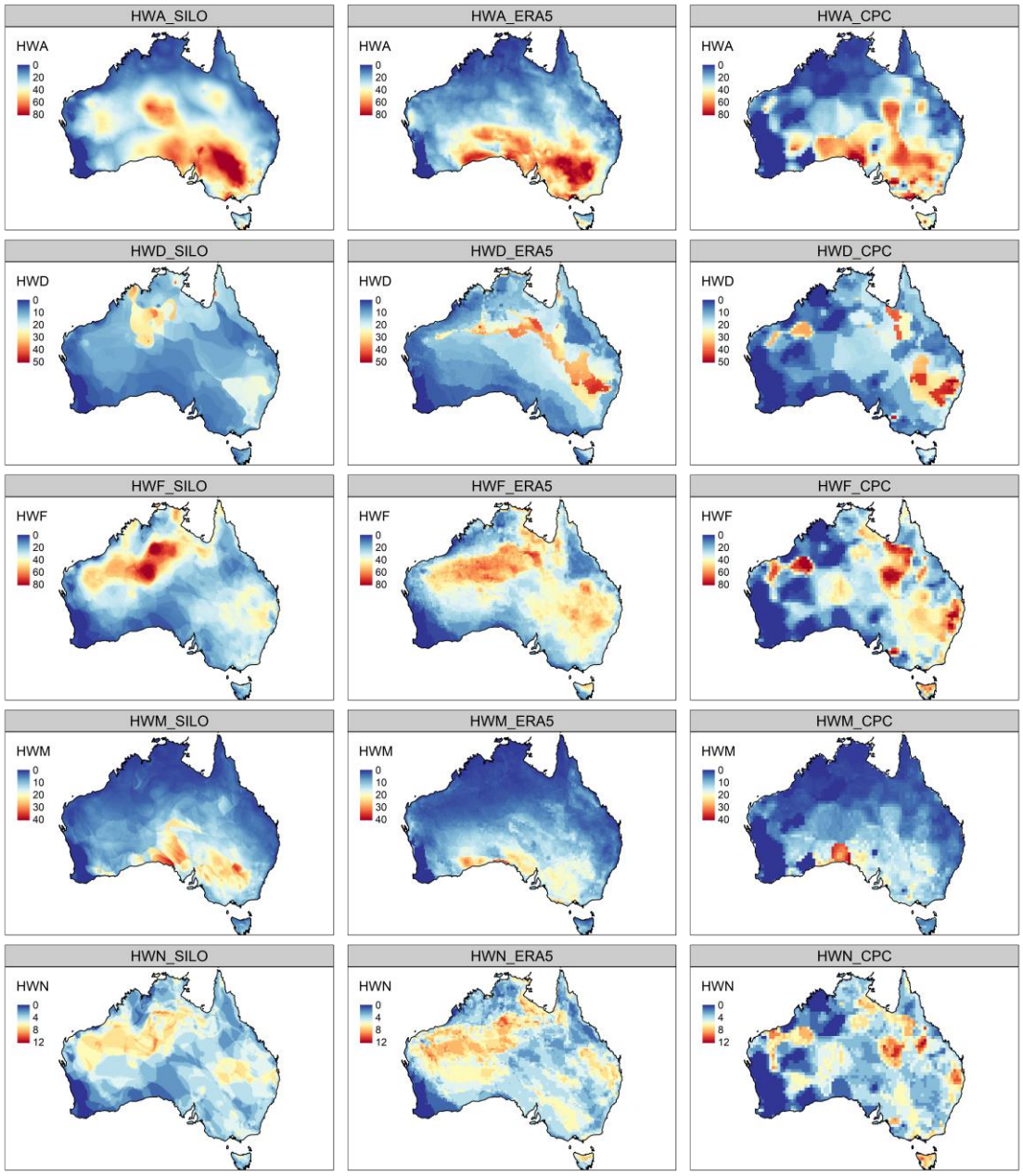


Figure 4-3 Examples of heat wave aspects derived from three different climate datasets in 2018

Table 4-2 Structural similarity index between different heat wave characteristics from three climate datasets.

Heat Wave Characteristics	ERA5_SILO	ERA5_CPC	SILO_CPC
Global SSI_{HWA}	0.78	0.68	0.67
Global SSI_{HWD}	0.77	0.67	0.66
Global SSI_{HWF}	0.71	0.59	0.58

Global SSI_{HWM}	0.76	0.74	0.69
Global SSI_{HWN}	0.59	0.56	0.55

4.3.3 How do the methods differ in identifying and characterising heat waves?

Five heat wave metrics for each method here are defined by ERA5 (seen in Figure 4-3). HWA measured by EHF ($°C^2$) tend to be higher than HWA (Tmax, °C) and HWA (Tmin, °C) due to the different units. Regions that display the higher values in HWA (Tmax) and HWA (Tmin) are very similar, mostly located in the southeast and central Australia. While the EHF-based HWA not only shows higher values in the southeast but also along the coastal regions of South Australia and Victoria. The extreme HWA by EHF all exists in the southward of 20°S. In contrast, HWA is not as large as expected in the northern tropical area. As HWM and HWA are related to heat wave intensity, their spatial patterns are largely similar. For those heat wave aspects (HWD, HWF) related to longevity in different ways, HWD and HWF defined by Tmax and Tmin are similar in spatial structure, which are centered in northwestern Australia and in eastern Australia. However, the lengths of HWD and HWF from Tmax and Tmin are about two times higher than HWD and HWF from EHF. Compared to northwestern Australia, HWF (EHF) is shorter at 60 days. Conversely to HWD and HWF, HWN produces different results in northwestern and eastern Australia where there are larger HWN variations from the EHF method. Figure 4-4 shows that the EHF based method identify four distinct heat wave events, while TX90 based method detects nine heat wave events and TN90 based method finds three heat wave events. The EHF method can combine the characteristics of both TX90 and TN90.

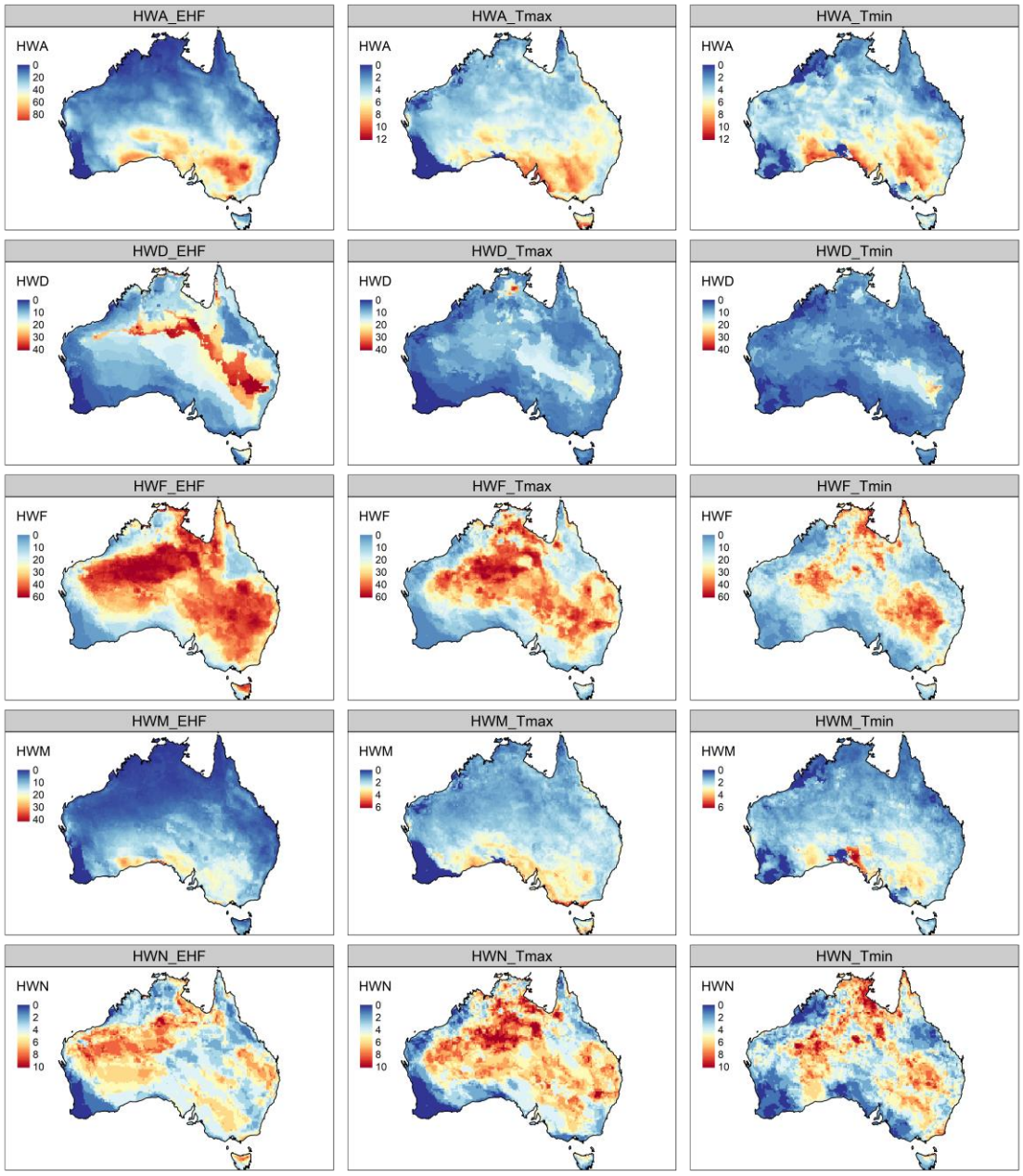


Figure 4-4 Examples of heat wave aspects of ERA5 from three different methods in 2018

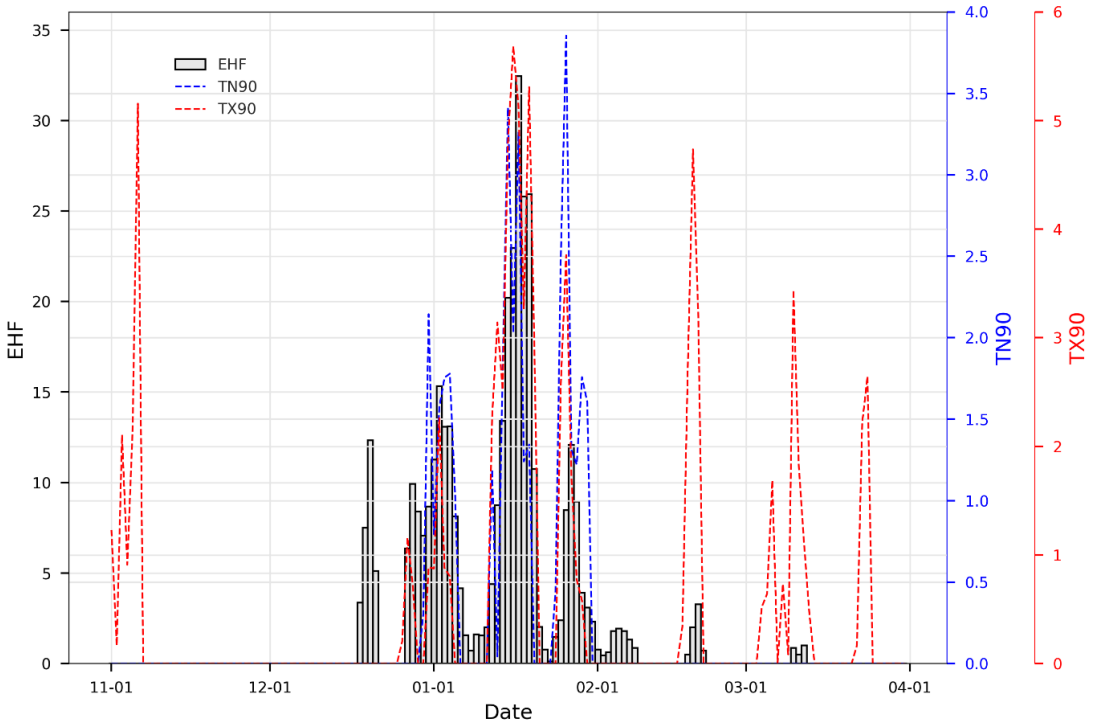


Figure 4-5 Distinct heat wave events derived from time series with EHF, TN90 and TX90 at the same point of southeastern Australia.

4.3.4 How does the heat wave risk change in recent climates?

To explore the heat wave risk in recent climates, the average values of HWA (the highest value of the heat wave in a season) over Australia for the past 100 years were used. Non-stationary return levels based on HWA versus the time covariate across the whole continent are generated by *NEVA*. As shown in Figure 4-5(a), the effective return levels vary over time indicating return level should be chosen for years to have the same probability of occurrence. For example, the effective return level (HWA) corresponding to a 25-year event during 1920-1944 is 37°C^2; the effective return level for a once-in-50-year event (1920-1969) should be 45° C^2 and the effective return level for a 100-year period (1920-2019) is 60°C^2. In Figure 4-5(a), we also observe that there is a strong upward trend ($p < 0.005$) for HWA over Australia during the 1920-2019 period. This suggests that heat wave amplitude was increasing under climate change. Figure 4-5(b) compares the probability density functions (PDF) of the HWA under two different time intervals (1920-2020, 1980-2020). We find that there is an obvious warming shift

of PDFs of the HWA during 1920-2020 compared with that during 1980-2020. This is consistent with the observed increasing trend in 4-5(a). In addition, the warm tail of the PDFs for the period of 1980-2019 is greater than that of 1920-2019 implying that extreme heat events have much higher probability with effects of climate change. We also find that the 2019 heat wave event is not rare (> 10-year effective return levels, 4-5(a)), with the PDF observed in 2019 for the 1980-2020 higher than that for the 1920-2020 as shown in Figure 4-5(b). From the long-term (1920-2019) and the short-term (1980-2019) time series of HWA, GEV fits are estimated together with the corresponding ±1.96 standard error for a 95% confidence interval in Figure 4-5(c), it denotes that the 2019 heat wave (HWA is 45.6 $^{\circ}C^{2}$) has a lower probability of occurrence over 1920-2020 climate and a higher probability over 1980-2020 climate (> 10-year return periods for GEV fit 1980-2020, Figure 4-5(c)).

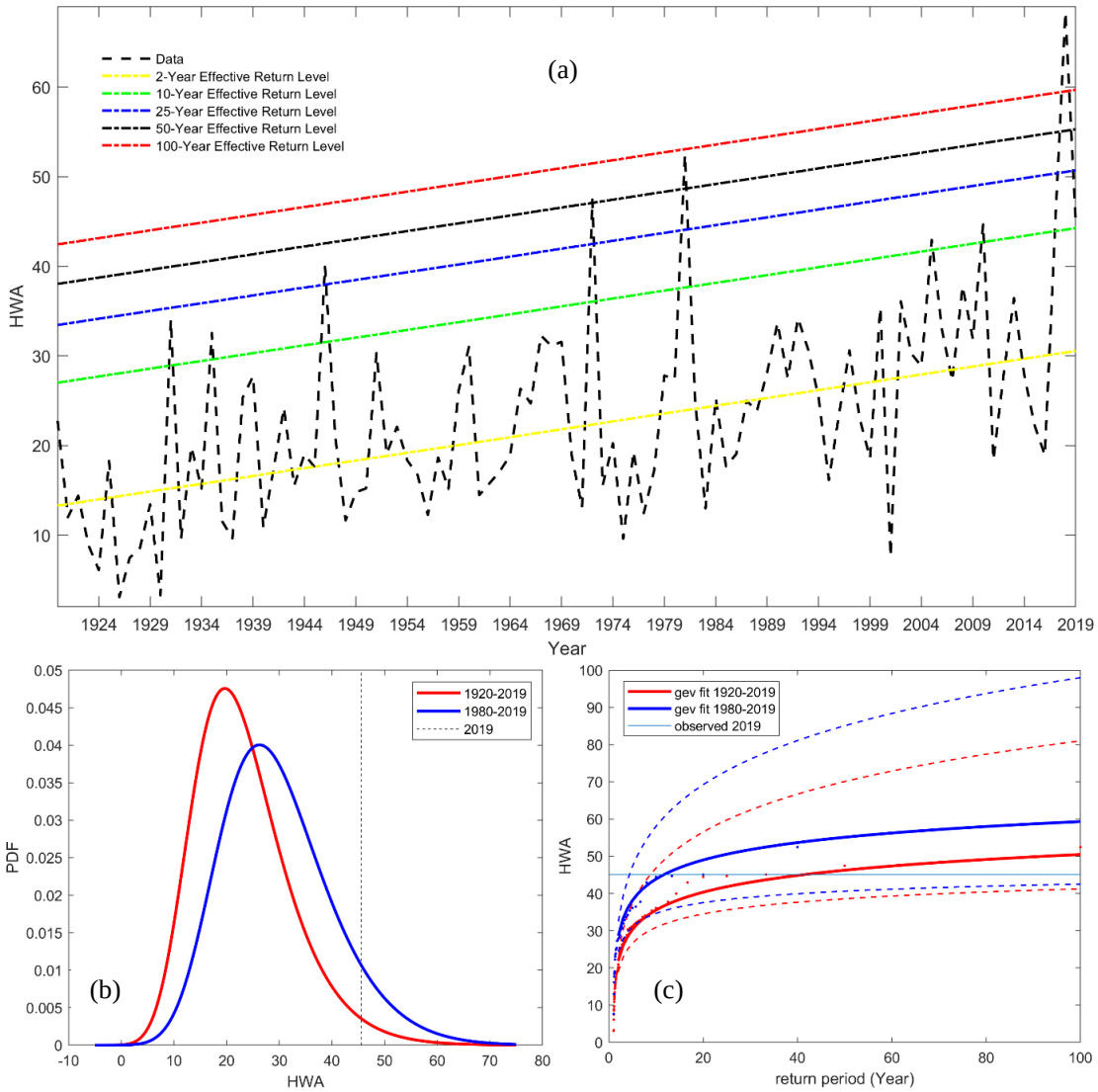


Figure 4-6 (a) Effective return level under the non-stationary assumption with mean HWA value from the continental Australia. (b) The probability density functions (PDF) of HWA under 1920-2019 and 1980-2019. (c) Return period of HWA over Australia. The distributions are fit with non-stationary GEV for the climates of 1920-2019 (red), 1980-2019 (blue).

4.3.5 How does the heat wave risk change under future climate conditions?

Figure 4-6 shows the near-future (2030-2060) and far-future (2069-2099) projected HWA using CMIP5 GCM datasets under two emissions scenarios compared with the 1976-2006 climate. Overall, HWA is projected to increase significantly during the two future periods and a larger fraction of southern Australia is projected to experience more extreme heat wave events.

We also see that the average HWA derived from CMIP5 multi-GCM ensemble mean ranges from 0-10 °C^2, and HWA decreases equatorward to ~3 °C^2 in the northern Australia. Under the two future periods of RCP4.5, the spatial extent of HWA mainly aggregates in the southern Australia. Compared with HWA in the near-future, HWA in the far-future expands from southeast to western and central Australia. Under the two future periods of RCP8.5, HWA not only increases its intensity but also expands from south to north. As expected, the change in HWA from RCP8.5 is more extreme than that from RCP4.5, indicating that greenhouse warming strongly amplifies the amplitude of heat wave events. Figure 4-7 shows the characteristic of HWD changes in the two future periods with different emission scenarios. The patterns of change for HWD are opposite to the change for HWA; northern Australia shows significant increases and southern Australia experience a moderate increase. In the far-future period of RCP4.5, we also note that HWD shows a stronger increase in western coastal areas and in northern tropical Australia, with HWD across northern tropical area reaching ~120 days. Again, in the far-future period of RCP8.5, HWD represents an amplification of the RCP4.5 pattern, that is, the duration of heat waves is much stronger than for RCP4.5. This indicates that the duration of southern Australia heat waves is not as sensitive to warming as those in northern Australia, largely due to the southern regions being associated with anticyclones and cold fronts.

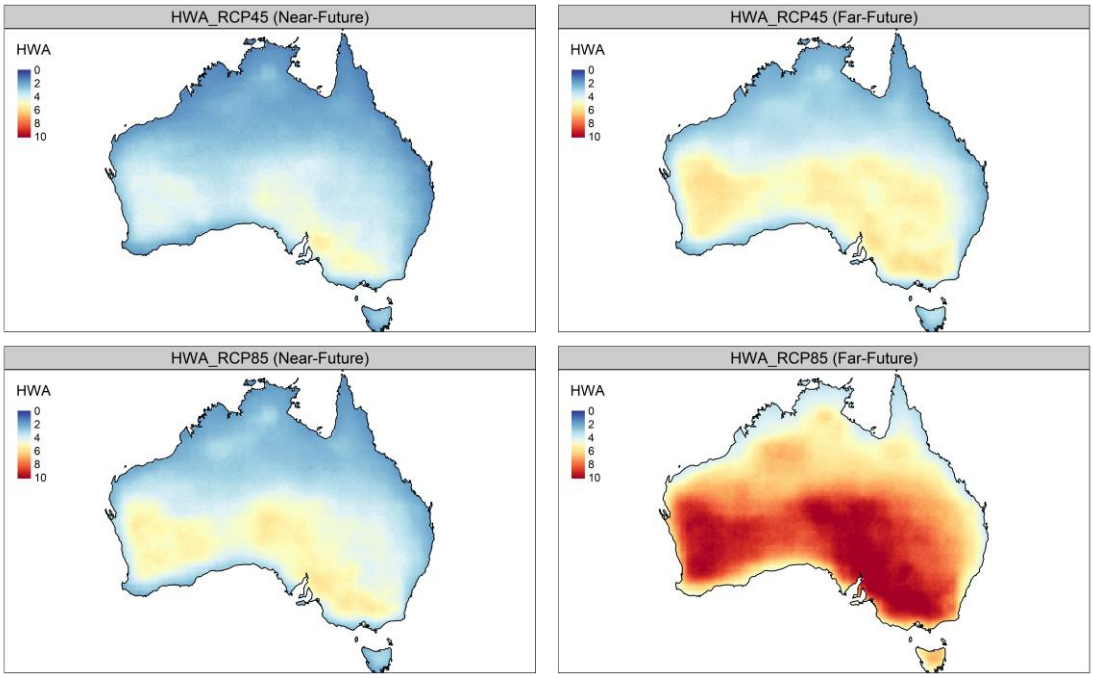


Figure 4-7 Near-future (2020–2039) and Far-future (2069-2099) projected climatology for heat wave amplitude obtained from the CMIP5 multi-GCM ensemble

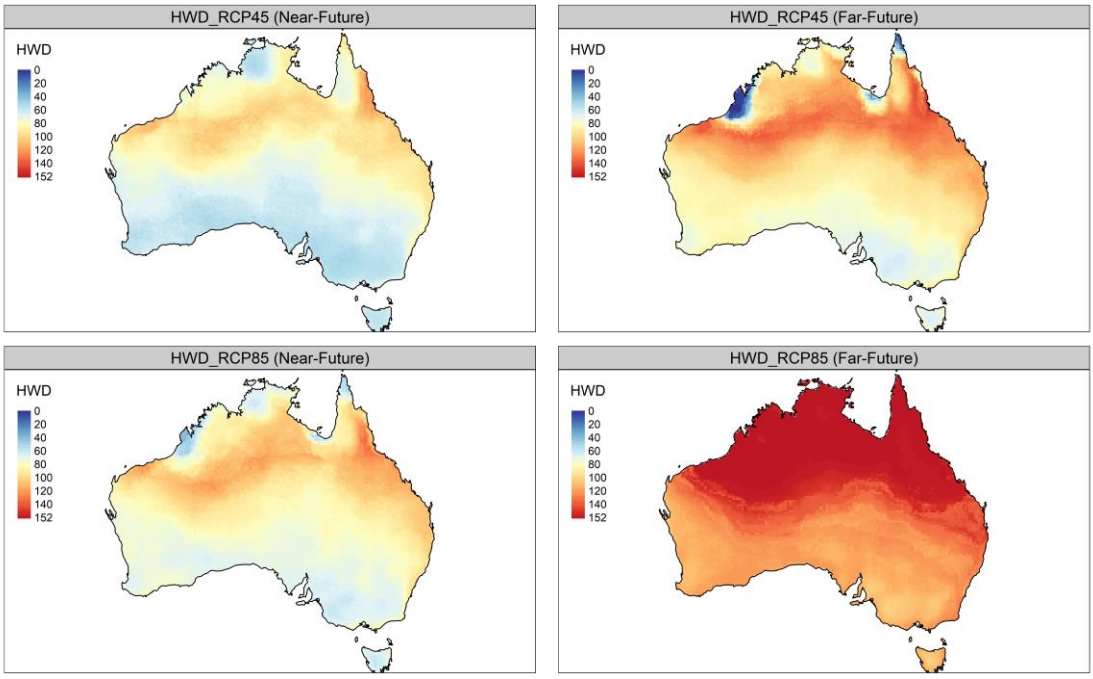


Figure 4-8 Near-future (2020–2039) and Far-future (2069-2099) projected climatology for heat wave duration obtained from the CMIP5 multi-GCM ensemble

4.4 Discussion

4.4.1 Model Comparison

To evaluate the performance of our model, we made a comparison with GHWR toolbox (https://github.com/mojtabasadegh/Global_Heatwave_and_Warm_Spell_Toolbox). For the comparison, the CPC datasets during a period of 1979 to 2019 were used to model heat wave metrics. Both software toolboxes apply EHF-based method to measure the heat wave metrics. Note that the definition of EHF is composed of the previous three-day mean and the previous thirty-day mean. The threshold of the 95th percentile of Tmean was calculated based on the 20 years period (1979 to 2009). Two 2018 heat wave indices are obtained from two different software packages (Figure 4-8). We can see that the spatial pattern of HWD from our model is consistent with that of GHWR. However, the comparison of HWM shows large difference in spatial patterns. Based on the HWM results of Alexander and Perkins (2013), the HWM of the northern Australia are no more than 12, as the tropical climate imposes less diurnal and seasonal variation in temperature than that in southern Australia. In contrast, the higher HWM values tends to occur in southern Australia and experience higher average peak values. Argüeso et al. (2015) reported higher HWM values towards the south-west of NSW and lower HWM values to the north coast of NSW, is consistent with the spatial pattern from our model. We also note that the HWM from GHWR (3-day average) has a similar spatial pattern similar to that of HWM from Argüeso et al. (2015), i.e., the highest values of HWM are found in the north-west corner and the lowest values in the mountains of the south. It means that the heat wave metrics from our model are consistent with the original definition of Alexander and Perkins (2013).

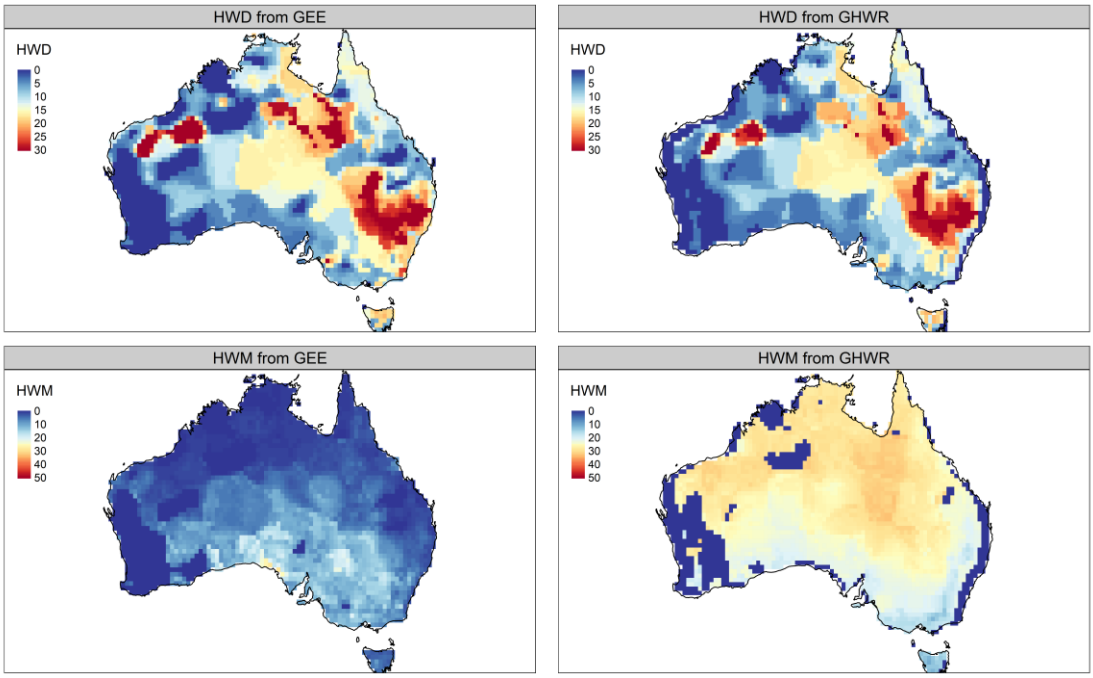


Figure 4-9 Heat wave metrics comparison between HWT and GHWR software tools

4.4.2 Heat wave threshold

The CMIP5 multi-GCM ensemble mean projects that longer summer heat waves will occur in northern Australia and hotter heat wave events will increase for southern Australia in the late twenty-first century, with more extreme change in the higher emission scenario RCP8.5 than for the lower emission scenario RCP4.5. The results reveal that the hottest heat waves will increase in southern Australia, which may account for the increasing trend of severe summer bushfires occurring in southeast Australia. Despite the different heat wave definitions and 25-member ensemble mean, our model results are consistent with the results from Purich et al. (2014). However, possibly due to the coarse resolution of the HWD from Purich et al. (2014), trends over Tasmania (an island state) are opposite to the overall pattern of change. While the patterns of change in Tasmania are consistent with the changes in other continental states, it means that our HWD results show promise in simulating fine-scale projections without using downscaling techniques.

Future extreme heat waves in our study are defined relative to a historical reference period,

we find a substantial increase in amplitude, duration and extent in both near-future and far-future periods (seen in Fig.6, Fig.7). For example, the duration of heat waves can even last over the entire warm season in some areas, which amounts to 152 days. The amplitude of heat waves significantly increases over southern Australia. Such results are not surprising and are in line with other findings (Lyon et al., 2019; Perkins-Kirkpatrick and Gibson, 2017). However, the sensitivity of heat waves to different heat wave thresholds was not explored. Vogel et al. (2020) identified future heat waves with different heat wave thresholds: fixed, seasonally moving and fully moving, where fixed thresholds are based on hot days relative to a historical baseline; seasonal and fully moving thresholds are defined by hot days relative to future conditions. They find that using fixed thresholds might overestimate future heat waves, while using seasonal and fully moving threshold results in little or no changes in future heat wave metrics. To better estimate heat wave characteristics and risk in a warming world, it would be useful to adopt varying heat wave thresholds for future spatiotemporal heat wave studies.

4.4.3 Future needs

For this study, we use the 5-km SILO gridded climate data, reanalysed data (25 km, 50 km) to estimate the heat wave at a large scale. However, those climate datasets do not take into account the smaller scale temperature variations, that is, the weather stations used to produce the gridded climate data were too sparse to record fine scale variations in extreme temperatures. For example, we find that the gridded climate data have relatively coarse spatial resolutions and cannot meet the need of monitoring heat wave variances in complex settings, and the heat wave maps are generally distributed evenly over urban heat islands. Furthermore, the location of most weather stations is away from building areas and the associated heat islands where extreme heat waves pose the greatest risk to human health. This issue can be at least partially resolved by using satellite thermal infrared sensing method to monitor and analyse heat waves at a local scale.

The proliferation of land surface temperature (LST) products offers an opportunity to study the characteristics of extreme heat waves at the community scale and give insight into urban

heat wave planning and the prevention of heat-related mortality. For example, MODIS LSTs have higher spatial resolution (1 km) and temporal resolution (four passes per day). MODIS LSTs provide the maximum and minimum products for the 20 years back to March 5th, 2000, which could be a valuable resource to capture extreme heat waves and for regional and local scale heat wave research. However, it is difficult to map LST accurately as the temperature are very variable and could be affected by climate factors like clouds and wind (Venter et al., 2020).

Compared to daily satellite data from MODIS (four passes a day), Himwari-8 data provides real time data at 10-minute intervals, but the spatial resolution is 2 km which is also suitable to conduct regional studies. The high temporal resolution of Himawari-8 can show the diurnal characteristics of extreme heat waves on urban heat waves. Despite the limitations of the relatively short time period (from 2015 to present) of the historical data archive of Himawari-8, a combination of MODIS LST and Himawari-8 LST offers a better solution for obtaining a higher spatial resolution while maintaining a higher temporal resolution, which is extremely useful for characterizing the heat wave characteristics and investigating the relationship between heat waves, land cover and population.

The health or agriculture impacts of heat waves are not only related with temperature measurements, but also affected by some additional factors. For example, health effects are associated with factors including perceived temperature, solar radiation, relative humidity, wind, while for agriculture, the parallel occurrence of droughts is highly relevant. Due to the problems of short time spans, inconsistency, and biases, these additional measurements have limitations on precisely capturing spatio-temporal pattern of heat wave impacts. The reason why we choose temperature-based heat wave definition is because it can be calculated from readily available climatological data and provides information on various aspects of heat waves. In other words, the choice of temperature rather than other measures is based on their feasibility across varying climates on long-term scales. Further, the availability of long-term temperature datasets at finer spatial scales can greatly improve our understanding of heat wave. We concur that the heat wave definitions directly rely on the critical temperature thresholds. However, there is no

universal temperature threshold for health impacts because of regional variability of health status, socio-economic factors, and demographic factors (Alexander and Perkins, 2013). This impact also exists in agriculture due to varying regional patterns of plant species and physiology. Therefore, a given threshold suitable in a small region may not be applicable to a continental study like ours. Fischer and Schär (2010) explored health-related heat wave indices in three health factors: heat wave duration, minimum temperature, and relative humidity. Our study also quantified the heat wave duration, minimum temperature-based heat wave indices. A combined calculation of temperature and humidity will be considered in our future study.

4.5 Conclusion

We have developed a heat wave toolbox that has the ability to estimate past, current and future changes in heat waves at a continental scale. It uses a well-known heat wave framework constructed by Alexander and Perkins (2013) and considers intensity, frequency, magnitude, duration and areal extent to explore the spatio-temporal evolution of heat wave severity and coverage. This study is the first attempt to estimate heat wave events across Australia using high spatio-temporal climate datasets. With these heat wave aspects from multi-source data and different methods, we were able to investigate the effects of scales, data quality and definition. We find that ERA5 datasets are the best in characterizing the heat wave events. In exploring the role of different methods on the identification of heat waves, we find that heatwave characteristics based on the Excess Heat Factor index provide more details on heatwave changes.

With the past 100 years of heat wave datasets, the HWA average mean values were calculated and used to estimate non-stationary return levels and return periods. We find that extreme heat wave events have much higher probability due to the effects of climate change. The heat wave event in 2019 may be more frequent in the coming decades. For the climate by the end of century, using heat wave metrics derived from a multi-model ensemble mean, we predict HWA to increase significantly during the two future periods and a larger fraction of southern Australia is projected to experience more extreme heat wave events. Furthermore, the

patterns of change for HWD are opposite to those for HWA; northern Australia shows significant increases and southern Australia experience a moderate increase. The methodology and the cloud computing-based toolbox (HWT) is useful for dynamic visualization, extraction, and processing of complex heat wave events, and applicable anywhere in the world.

Chapter 5. New assessment of water and wind erosion for Australia 2000-2020

This chapter is based on the following manuscript:

Zhang, M., Yang, X., Leys, J., Gray J., Zhu, Q., Yu, Q. (2021). The first combined water and wind erosion assessment for Australia ¬2000-2020 (Ready submission for Catena).

Abstract

Soil erosion caused by water and wind is a complex natural process that has been accelerated by human activity. This erosion has resulted in increasing areas of land degradation which threaten the productive potential of landscapes. Consistent and continuous erosion monitoring will help identify the trends, magnitude, and location of soil erosion. This information can then be used to evaluate the impact of land management practices and inform programs that aim to improve soil condition. We apply the Revised Universal Soil Loss Equation (RUSLE), Revised Wind Erosion Equation (RWEQ), and an albedo-based wind erosion model to simulate water and wind erosion dynamics. With the advent of new or improved earth observation big data, monthly and annual water, and wind erosion estimates at high spatial resolution are produced for Australia from 2000 to 2020. We also evaluate the performance of three gridded precipitation products for rainfall erosivity estimates using ground-based rainfall. For model validation, water erosion products are compared with existing products and wind erosion results are also compared with other models.

Key words: *Soil erosion, water and wind erosion, Earth Observation, Google Earth Engine*

5.1 Introduction

Soil erosion is a major threat to sustainability of agriculture (Borrelli et al., 2017; FAO, 2015). Under changing land use and climate, soil erosion from water and wind has accelerated with resulting economic, social, and environmental implications, both on-site and off-site (Telles et al., 2013). On-site, water and wind erosion causes the loss of soil, nutrients and organic matter that results in decreased soil fertility and land productivity (Zhang et al., 2019). The reduced productivity of farmland means that about 10 million ha of cropland worldwide is abandoned yearly due to soil erosion(Chappell et al., 2019; Faeth, 1994). This further leads to reduce the social viability and population levels in rural communities, influencing long-term sustainable regional development. The subsequent sedimentation and nutrient loss may also cause off-site environmental, air (Middleton, 2019) and water quality degradations.

In Australia, for example, the assessment Bui et al. (2010) concluded that soil erosion in Australian cropping regions was occurring at unsustainable rates and has a critical impact on agricultural productivity. Environmental impacts of excessive sedimentation and nutrient delivery on inland waters, estuaries and coasts are already occurring. The net median erosion rate in cultivated regions is estimated 1.26 Mg ha^{-1} yr^{-1} (Chappell et al., 2011), and 7% of Australia had soil losses of more than 1 Mg ha^{-1} yr^{-1}. It also should be noted that Australia is the most fire-prone regions of the world. Wildfire related water erosion in Australia is responsible for reef deterioration, roads damage, river pollutants (Yang et al., 2020). In addition, wind erosion from arid and semi-arid areas of Australia severely affects the air quality in the coastal zone where most Australians live (Leys et al., 2011). Since 2000, the millennium drought and mega-fires in Australia also prompt the urgent need to revisit soil erosion dynamics to provide a more contemporary view of water and wind erosion trends.

Over the last century, extensive studies have been conducted to estimate soil erosion using various monitoring and modelling approaches. From the perspective of representation of soil erosion, water and wind erosion models are classified into three groups viz. Empirical, Processed-based, and machine learning models (Jarrah et al., 2020; Karydas et al., 2014). Each

of these models has found its niche for different reasons. Empirical water erosion model like RUSLE is relatively easy to parameterize. Its simplicity enables the handling of large datasets and computation, making its attractive for large-scale land degradation assessments. Process-based water erosion models like Water Erosion Prediction Project (WEPP) and Griffith University Erosion System Template (GUEST) can gain insights to the space-time evolution of soil erosion process involved (Laflen et al., 1997; Yu and Rose, 1999). However, the non-linear and physical parameters of processed-based models need to be calibrated by many input observations, and it is hard to apply over large areas. In recent years, large and abundant Earth Observation (EO) datasets are becoming public available to scientists. Rather than using traditional models, big data-based machine-learning (ML) methods (like SVM, RF, ANN) have been successfully used for landslides, debris flows, and gully erosion (Rahmati et al., 2017). While water erosion models were initially developed in agricultural area, wind erosion models have mainly applied in drylands where dust originates. To monitor the wind erosion hazard at high spatio-temporal resolution, empirical models like RWEQ, and process-based models like Single-Event Wind Erosion Evaluation Program (SWEEP) and albedo-based model have been developed to assess the spatial and temporal patterns in erosion dynamics with numerous landscapes (Chappell and Webb, 2016; Fryrear et al., 2000; Tatarko et al., 2019). Commonly used RWEQ model has a capability to consider the impacts of weather, soil, vegetation, and roughness factors on the rate of soil loss (Zhang et al., 2019). Chappell et al. (2018) demonstrated that albedo-based approximation of aerodynamic sheltering could improve wind erosion estimation over large area.

Due to its simple linear equation form, the RUSLE water erosion model has been applied to estimate the potential global soil erosion under the land use and climate change (Borrelli et al., 2020). Similar to RUSLE, the limited need for observation data as well as the acceptable prediction performances makes RWEQ model an ideal tool for large-scale wind erosion prediction (Borrelli et al., 2016; Pi et al., 2017; Youssef et al., 2012). Teng et al. (2016) was among the first who assimilated the latest big EO datasets to improve the soil loss estimation

over the continent of Australia. For example, Teng et al. (2016) simply allocated cover factor (C) of RUSLE to each land cover type using the Dynamic Land Cover Dataset (DLCD) (Yang, 2020). McKenzie et al. (2017) recommend using fractional vegetation cover to replace the DLCD to estimate C factor across Australia. The big advantage of fractional cover is that is dynamic, changing monthly, and represents the outcome of land management practices. In addition, rainfall erosivity (R) in Teng's method is estimated using the Tropical Rainfall Measuring Mission (TRMM), however, Atiqul Islam et al. (2020) showed that the Global Precipitation Measurement (GPM) performed the best for satellite precipitation estimation in terms of local mean, variance, and covariance for Australia. Therefore, the comparison of different precipitation products for soil erosion modelling across Australia is still essential. Jiang et al. (2019) applied RUSLE and RWEQ model to assess the soil erosion dynamics in China's Loess Plateau. It is still a challenge to apply both water and wind erosion to large area and the applications of these commonly used models in predicting water and wind erosion across Australia has not been undertaken.

In this study, we aimed to i) parameterise the RUSLE, RWEQ model using the latest earth observation datasets (GPM, SRTM, MODIS Fractional Cover and Albedo products) and digital soil mapping methods; ii) evaluate rainfall erosivity from three different satellite precipitation products and validate with ground-based rainfall erosivity; iii) estimate monthly and annual soil loss by water and wind across Australia from 2000 to 2020; iv) assess and compare RWEQ-based wind erosion results with that from the albedo-based wind erosion model.

5.2 Data and methods

5.2.1 Earth Observation and Soil Datasets

In this study, EO datasets from multiple sources were collected to model water and wind erosion, Table 5-1 provides information about these datasets. Both the ground-based and reanalysis climate datasets were used. SILO is a database of Australian climate data from 1889 to the present hosted by the Queensland Department of Environment and Science (DES)

(https://www.longpaddock.qld.gov.au/silo/). It provides daily climate variables on a 0.05° grid across Australia for research, modelling and climate applications. Currently, SILO data are uploaded into GEE Data Catalogue and maintained by EO Data Science, available online (https://www.eodatascience.com/). In SILO, rainfall datasets have a spatial resolution of 5 km. Satellite-based rainfall datasets are also used for rainfall erosivity estimates. The TRMM 34B2 product contains TRMM-adjusted precipitation (mm/hr) and are available at 0.25° spatial resolution and 3-hourly temporal resolution. Compared with TRMM precipitation, GPM has the same temporal resolution as TRMM, but a different spatial resolution of 0.1° and is more sensitive to light rainfall. For reanalysis climate data used in this study, Global Land Data Assimilation System (GLDAS) is an important global-scale data source for land surface states and flux (e.g., soil moisture and wind speed), with the spatial resolution of 0.25° and the temporal resolution of 3-hourly (Rodell et al. (2004b). In addition, we used the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis 5 (ERA5) atmospheric reanalysis of the Australian climate to retrieve variables for RWEQ model. For more information on ERA5, readers can refer to Hersbach et al. (2020). In GLDAS and ERA5, both the soil moisture and wind speed variables have a spatial resolution of 0.25° and 3-hourly temporal frequency. Note that the albedo-based model of horizontal and vertical sediment flux (F_d) and RWEQ model (S_L) were applied across Australia from 2000 to 2020. Based on GLDAS datasets, the three-hourly wind and soil moisture data aggregated to daily data using the maximum value. The daily estimates of F_d and S_L were then aggregated to produce monthly mean F_d and S_L for the period 2000-2019.

To calculate the slope length and steepness and soil roughness factor of RUSLE and RWEQ model, hydrologically enforced Digital Elevation Model (DEM) are used at approximately 30 m horizontal resolution. The most recent fractional vegetation cover products are used to estimate the vegetation cover factor in RUSLE and RWEQ models (Guerschman and Hill, 2018). The lateral cover in the Albedo-based wind erosion model was estimated using the MCD43A1 V6 Bidirectional Reflectance Distribution Function and Albedo (BRDF/Albedo)

model parameters dataset, which is a 500-m daily product (https://lpdaac.usgs.gov/products/mcd43a1v006/). The soil attributes were obtained from the Soil and Landscape Grid of Australia (SLGA), a three-dimensional Australia soil grid including 11 soil attributes and confidence intervals at 3 arc second resolution (~90 m pixels). For further modelling details refer to Viscarra Rossel et al. (2015).

Table 5-1 Datasets used in this study

Dataset	Description	Spatial Resolution	Time Period	Data Source
SILO	Scientific Information for landowner's climate database	5000 m	1920-2020, daily	GEE
DEM	Australian SRTM Hydrologically Enforced Digital Elevation Model	90 m	2010	GEE
FVC	Fractional Vegetation Cover	500 m	2000-2020, daily	CSIRO (GEE)
Albedo	MODIS BRDF-Albedo Model Parameters	500 m	Daily	GEE
GPM	Global Precipitation Measurement	10000 m	3 hours	GEE
TRMM	The Tropical Rainfall Measuring Mission	25000 m	3 hours	GEE
SLGA	Soil and Landscape Grid of Australia	90 m	-	GEE

GLDAS	Global Land Data Assimilation System	25000 m	3 hours	GEE
ERA5	ECMWF Reanalysis 5 (ERA5) atmospheric reanalysis	25000 m	3 hours	GEE

5.2.2 Estimates of Water Erosion by RUSLE

The water erosion in Australia was estimated based on RUSLE model as the following equation:

$$A = R \times K \times LS \times C \times P \quad (1)$$

where A is the predicted soil loss ($Mg \cdot ha^{-1} \cdot yr^{-1}$); R is the rainfall erosivity factor ($MJ \cdot mm \cdot ha^{-1} \cdot h^{-1} \cdot yr^{-1}$); K is the soil erodibility factor ($Mg \cdot h \cdot MJ^{-1} \cdot mm^{-1}$); LS is the slope length and steepness factor (dimensionless); C is the cover and management factor (0-1, dimensionless); and P is the conservation support-practices factor (0-1, dimensionless). P factor is set to be as a constant one in this study for reasons that water erosion control practices in Australia are extremely rare.

5.2.2.1 Rainfall erosivity (R) factor

Among these factors, rainfall erodibility is the most important input parameter in RUSLE model to describe water erosion dynamics. We used a regional daily rainfall erosivity model Yang and Yu (2015) and idealized intensity distributions (Brown and Foster, 1987) to estimate the R factor and compared three different rainfall data sources including Bureau of Meteorology (BoM) gridded daily rainfall, satellite precipitation (TRMM and GPM). The multiple data sources provide means of cross comparison and validation.

$$EI_j = \alpha[1 + \eta \cos(2\pi fj - \omega)] \sum_{d=1}^{N} R_d^{\beta} \quad (2)$$

$$\mathrm{E} = \sum_{r=1}^{N} e_r \cdot \Delta V \cdot I \quad (3)$$

$$e_r = 0.29[1 - 0.72e^{-0.05I}] \quad (4)$$

where EI_j is the month j rainfall erosivity, R_d is the daily rainfall (mm), N is the rain days of the month j. The model parameters (α, β, η) are modelling coefficients. e_r is the unit rainfall energy, I is the rainfall intensity, ΔV is the rainfall volume, E is event erosivity.

5.2.2.2 Cover-management (C) factor

The cover and management (C) factor was derived from the most recent (Version 3.1.0) and validated fractional vegetation cover products including both photosynthetic (PV) and non-photosynthetic (NPV) vegetation (Guerschman and Hill, 2018; Guerschman et al., 2015) with a spatial resolution of 500 m. The monthly C factor time series for all Australia continent were produced e following the method as described in Yang (2014).

$$C_j = \mathrm{e}^{\left(-0.799-7.74\times GC_j+0.0449\times GC_j^2\right)} EI_j/EI_t \quad (5)$$

where C_j = RUSLE cover and management factor in month j (1–12), GC_j = groundcover (0–1) in month j (1–12), EI_j is the erosivity index (EI) or rainfall erosivity in month j, and EI_t is the total rainfall erosivity or the R factor in that year. Ancillary data (i.e., NDVI) and GIS mask layers for water bodies, snow cover, rocky surfaces, cropping, and urban areas were applied to readjust the C factor values and fill in the no-data gaps (Yang, 2014). The C factor was resampled to 90 m resolution so that to match with the resolutions of other RUSLE factors.

5.2.2.3 Slope-steepness (LS) factor

The slope-length and steepness (LS) factor (unitless) was calculated from hydrologically corrected digital elevation model (DEM-H) based on comprehensive algorithms considering cumulative overland flow length (Yang, 2015). The 1-second DEM-H was resampled to ~90 m and used for the LS factor calculation based on Australia natural resource management (NRM) regions, then merged them to form a seamless LS-factor digital map at a spatial resolution ~90m for Australia continent. We chose the 1-second DEM-H because it was the best available DEM data across Australia with hydrological correction (Wilson et al. 2011). Automated scripts have been developed and implemented in a geographic information system (GIS) for the LS calculation. Compared with the LS values calculated using the 1-second DEM-H over NSW (Shan et al., 2019), there was an overestimate in the slope-length (L) sub-factor but an underestimate in the steepness (S) subfactor resulting an overall coefficient of efficiency about 0.70. It is the LS factor map with the highest spatial resolution in Australia so far.

5.2.2.4 Soil erodibility (K) factor

The soil erodibility factor RUSLE-K was estimated based on the RUSLE:

$$K = [2.1M^{1.14}(10^{-4})(12 - OM) + 3.25(2 - SS) + 2.5(PP - 3)]/100 \times 0.1317 \quad (6)$$

where OM is percent soil organic matter (= soil organic carbon ×1.72); SS is the soil structure code; PP is the soil profile permeability class; M is the particle size parameter that defines the relationship between percentages of silt, very fine sand (VFS) and clay content:

$$M = (\text{Silt} + \text{VFS}) \times (100 - \text{Clay}) \quad (7)$$

where Silt is % silt content (0.002–0.05 mm); VFS is % very fine sand content (0.05–0.1 mm); and Clay is % clay content (< 0.002 mm) based on USDA classification. The soil texture and organic matter data were from the Soil and Landscape Grid of Australia (Viscarra Rossel et al., 2015) with a spatial resolution of 90 m. Soil structure and permeability data were derived from the saturated hydraulic conductivity and grade of pedality attributes of the Australian Soil Resource Information System (McKenzie et al., 2000). The estimated K values range between 0.01 to 0.11 Mg·ha·h·ha^{-1}·MJ^{-1}·mm^{-1} with more variation compared to the estimates (0.02 to 0.04 Mg·ha·h·ha^{-1}·MJ^{-1}·mm^{-1}) from a previous study in Australia (Teng et al., 2016). The estimated K factor values agree well with a recent study in NSW (Yang et al., 2018). The K factor map has a spatial resolution of 90 m which matches with the Slope-steepness (LS) factor.

5.2.3 Estimates of Wind Erosion by Revised Wind Erosion Equation (RWEQ)

RWEQ model includes weather factor (WF, kg·m^{-1}), erodibility factor (EF, dimensionless), soil crust factor (SCF, dimensionless), roughness (K, dimensionless), vegetation factor (C, dimensionless). The horizontal sediment flux (Q in kg·m^{-1}) and vertical sediment flux (S_L) are estimated with the following the equations:

$$Q = Q_{max}\left(1 - e^{(Z/S)^2}\right) \tag{8}$$

$$S_L = \frac{2 \cdot Z}{S^2} Q_{max} \cdot e^{-(Z/S)^2} \tag{9}$$

where Q_{max} is the maximum transport capacity (in kg·m^{-1}), Z is the calculated distance of downwind wind (m), and S is the critical field length (m).

$$Q_{max} = 109.8\ [WF \cdot EF \cdot SCF \cdot K \cdot C] \tag{10}$$

$$S = 150.71(WF \cdot EF \cdot SCF \cdot K \cdot C)^{-0.3711} \tag{11}$$

The WF integrates the effects of various meteorological factors on wind erosion. It is calculated as:

$$WF = W_f \cdot \frac{\rho}{g} \cdot SW_f \cdot SD \tag{12}$$

$$W_f = \sum_{i=1}^{i-n} U_2\ (U_2 - U_t)^2\ \frac{N_d}{500} \tag{13}$$

$$SW_f = 1 - SW \quad (14)$$

$$SD = 1 - \mathrm{P} \quad (15)$$

where W_f is the wind factor (kg·m^{-1}), ρ is air density, g is gravity acceleration, SW_f is the soil moisture factor, SD is the snow cover factor and P is the probability that the snow cover depth is greater than 25.4 mm in the simulation period. U_2 is the wind velocity at 2 m; U_t is the wind velocity threshold, generally set as 5 m·s^{-1}; N_d the number of days in the simulation period (1 day in this study).

The soil erodibility factor (EF) and soil crust factor (SCF) are expressed by the aggregation of soil particles (especially clay, silty, and soil organic carbon) as follows:

$$EF = \frac{29.09+0.31S_a+0.17S_i+\frac{0.33S_a}{C_l}-2.59(SOC*1.72)-0.95CaCO_3}{100} \quad (16)$$

$$SCF = \frac{1}{1+0.0066(C_l)^2+0.021(SOC*1.72)^2} \quad (17)$$

where, Sa is the sand grain proportion; Si is the soil silt proportion; Sa/C_l is the ratio of soil sand grain and clay; C_l is clay proportion; SOC is soil organic carbon proportion; and CaCO3 is calcium carbonate proportion.

The calculation of surface roughness factor (K') is based on random roughness factor (C_{rr}) and soil roughness (K_r). As soil roughness (k_r) is difficult to estimate, the topographic roughness (K_r) is used instead of soil roughness. It is calculated as follows:

$$K' = e^{\left(1.86k_r-2.41{k_r}^{-0.934}-0.127C_{rr}\right)} \quad (18)$$

$$K_r = 0.2 \times \frac{\Delta H^2}{L} \quad (19)$$

where L is the topographic fluctuate parameters; and ΔH is the difference of elevation within distance L.

The vegetation cover factor is shown as follows:

$$C = e^{-0.0438FVC} \quad (20)$$

where, FVC is fractional vegetation cover (in %).

5.2.3 Albedo-based wind erosion model

Kok et al. (2014) describes a physically based wind erosion model. A critical approximation in wind erosion modelling is that momentum extracted by roughness elements which can be expressed by roughness density lateral cover. The lateral cover underpins current wind erosion models. However, the estimation of lateral cover over large areas is challenging and often approximated by remote sensing vegetation cover. By replacing the use of lateral cover which closely relates to sheltered area, Chappell and Webb (2016) provided an alternative parameterisation which equated sheltering to shadow and enabled the use of remote sensing albedo data. This albedo-based approach to horizontal and vertical flux modelling (Chappell et al., 2018), is considerably simplified without losing information content, thereby reducing uncertainty, and improving the wind erosion modelling used by Kok et al. (2014). Below is a summary of the horizontal sediment flux (Q_h, g·m^{-1}·s^{-1}) and vertical sediment flux (F_d, kg·m^{-2}·s^{-1}) from the MODIS albedo-based wind erosion model.

$$Q_h = c_{shao}\frac{\rho_a u_{S*}^3}{g}\left(1-\left(\frac{u_{*ts}(D)\mathrm{H}(w)}{u_{S*}}\right)^2\right) \quad (21)$$

$$F_d = C_d f_{bare} f_{clay}\frac{\rho_a(u_{S*}^2-u_{*t}^2)}{u_{*st}}\left(\frac{u_{S*}}{u_{*t}}\right)^{C_\alpha\frac{u_{*st}-u_{*st0}}{u_{*st0}}}, (u_{S*}>u_{*t}) \quad (22)$$

where $C_{shao}=0.006$, $C_\alpha=2.7$, $C_e=2.0$, $C_{d0}=4.4\times10^{-5}$, $\rho_a=1.23$ kg·m^{-3} and $\rho_{a0}=1.225$ kg·m^{-3}. The standardised threshold friction velocity of an optimally erodible soil $u_{*st0}=0.16$ m·s^{-1}. C_d is calculated in Equation. The fclay is the clay decimal fraction in the surface soil. The fbare can be estimated from the inverse of u_{S*} with a maximum value of 0.6 m·s^{-1} for Australia.

$$C_d = C_{d0}e^{\left(-c_e\frac{u_{*st}-u_{*st0}}{u_{*st0}}\right)} \quad (23)$$

$$f_{bare} = \max(u_{S*})/u_{S*} \quad (24)$$

The soil threshold friction velocity standardised to an atmospheric density (m·s^{-1}) as follows:

$$u_{*st} = u_{*t}\sqrt{\rho_a/\rho_{a0}} \quad (25)$$

where $\rho_a = 1.23$ kg·m^{-3} and $\rho_{a0} = 1.225$ kg·m^{-3}.

The main variables include the albedo-based soil friction velocity u_{S*} (m·s^{-1}), soil threshold friction velocity u_{*t} (m·s^{-1}) and $H(w)$. To calculate the surface soil friction velocity (u_{S*}) for a given freestream velocity of a particular pixel using the rescaled and normalised shadow (ω_{ns}):

$$\frac{u_{S*}}{U_f} = 0.0306\left(\frac{e^{-\omega_{ns}}{}^{1.1202}}{0.0125}\right) + 0.0072 \quad (26)$$

where Uf is the freestream wind speed at a height of 10 m. We invert MODIS Black-Sky Albedo data ω to estimate shadow which is then normalised by dividing it by BRDF parameter ω_0 to remove the spectral influences and then rescaled (ω_{ns}) for use with the calibration functions. When ω_{ns} is calculated, it is inserted into Equation to calculate u_{S*}.

$$\omega_n = \frac{1-\omega}{\omega_0} \quad (27)$$

$$\omega_{ns} = \frac{(a-b)(\omega_n - 35)}{(0-35)} + b \quad (28)$$

The $u_{*t}(D)$ is calculated using the bare soil threshold friction velocity $u_{*ts}(D)$ for a given size fraction D and the function of surface soil moisture $H(w)$ (dimensionless). Function $H(w)$ accounts for the volumetric soil moisture content w (m^3·m^{-3}) based on the difference between the potential w' based on clay content and near surface w:

$$H(w) = \sqrt{1 + (1.21(w' - w)^{0.68})} \quad (29)$$

$$w' = 0.0014clay^2 + 0.17clay \quad (30)$$

$$u_{*t}(D) = u_{*ts}(D)H(w) \quad (31)$$

$$u_{*ts}(D) = \left(A_N\left(\frac{\rho_p g D}{\rho_a} + \frac{\Gamma}{\rho_a D}\right)\right)^{-0.5} \quad (32)$$

where A_N = 0.0123 is a scaling coefficient, ρ_p = 2650 kg·m^3 is the particle density, g = 9.81 m·s^{-1} is the acceleration due to gravity, Γ = 0.000165 kg·s^{-2} is a parameter accounting for cohesive force.

5.2.4 DustWatch PM10 measurements

DustWatch has 39 measurement sites (denoted DWN) in south-east Australia. The Tibooburra DWN used in this study had an 8520 model DustTrak® from January 2009 to July 2019 and DRX model DustTrak®, from August 2019 to December 2020. DustTraks measure atmospheric aerosol concentration of PM10 as described in (Leys et al., 2018). During our study period, Tibooburra is selected as the monitoring sites against with model estimates because it is the dustiest site in NSW. In summary, hourly minute PM 10 concentration data is averaged to give monthly concentration (mg/m^3) and then converted to the vertical flux ($F_h, mg/m^3$):

$$F_h = PM10 * U_f \quad (33)$$

5.3 Results

5.3.1 Estimation of sub-factors in RUSLE

Maps of the water erosion factors in Australia are shown in Figure 5-1. The multi-year mean rainfall erosivity for Australia from 2001 to 2020 is show in Figure 5-1(a), with high rainfall erosivity in the coastal areas and tropics. The multi-year average R factor calculated from SILO rainfall datasets and ranges from 188 MJ·mm·ha^{-1}·h^{-1}·y^{-1} to 19,708 MJ·mm·ha^{-1}·h^{-1}·y^{-1}. TRMM-based R factor have values between 52 MJ·mm·ha^{-1}·h^{-1}·y^{-1} and 3,594 MJ·mm·ha^{-1}·h^{-1}·y^{-1} and GPM-based R factor between 187 MJ·mm·ha^{-1}·h^{-1}·y^{-1} to 25764 MJ·mm·ha^{-1}·h^{-1}·y^{-1}. We find that TRMM tends to underestimate the rainfall erosivity, while GPM-based R overestimates rainfall erosivity but is much closer to SILO-based calculation but has a larger maximum value.

In Figure 5-1(c), the mean C-factor map (2001-2020) derived from fractional vegetation cover products shows great spatial variation across Australia, reflecting the spatial distribution of vegetation and land cover types. Lower C factor values (dimensionless, less than 0.001) coincide spatially with high density vegetation cover along the Great Dividing Range. Higher C factor values (great than 0.04) are in the desert and arid areas where vegetation is very limited. C-factor values also show strong monthly temporal variations, more variation in the central

semi-arid and arid areas compared to coastal areas with high vegetation cover (Figure 5-2). Figure 5-1(d) shows the LS Factor layer with a cell size ~90m. LS factor in Australia continent has a great spatial variation within different landscapes. Higher values are in the south-eastern coastline and mountain regions, while smaller values are found mostly in the lowland of central Australia. The 90m LS factor in this study was compared with the reference LS factor values with 300 random points across NSW from the literature (Yang, 2015). In general, the quality of LS factor values would decrease as the DEM resolution become coarser. However, the 90m LS factor values in this study shows a good correlation (R^2=0.91). That is because as DEM resolution become coarser, the length sub-factor (L) becomes larger whilst slope sub-factor (S) tends to be smoother (Shan et al., 2019). Therefore, the LS factor of 90m in this study matches well with the referenced LS of 30 m under the combined influence of both L and S sub-factors.

Along the Great Dividing Range in south-eastern Australia, 60 sites were sampled to assess the rainfall erosivity derived from different data sources. Figure 5-3 shows the linear regression of rainfall erosivity from GPM and TRMM compared to SILO for each month. We find that GPM and TRMM have reasonably good fits in estimating the R factor over southeast Australia rainy seasons (spring and summer) but have poor performance in dry seasons. Comparing the monthly RMSE and R^2, TRMM-based R factor values are generally better than GPM-based R factor values with a systematic underestimation. When the rainfall erosivity is less than 300 MJ mm ha^{-1} h^{-1} m^{-1}, there is little correlation between satellite-based and SILO-based rainfall erosivity. This indicates that improved performance should be made to detect light rainfall in dry seasons (autumn and winter).

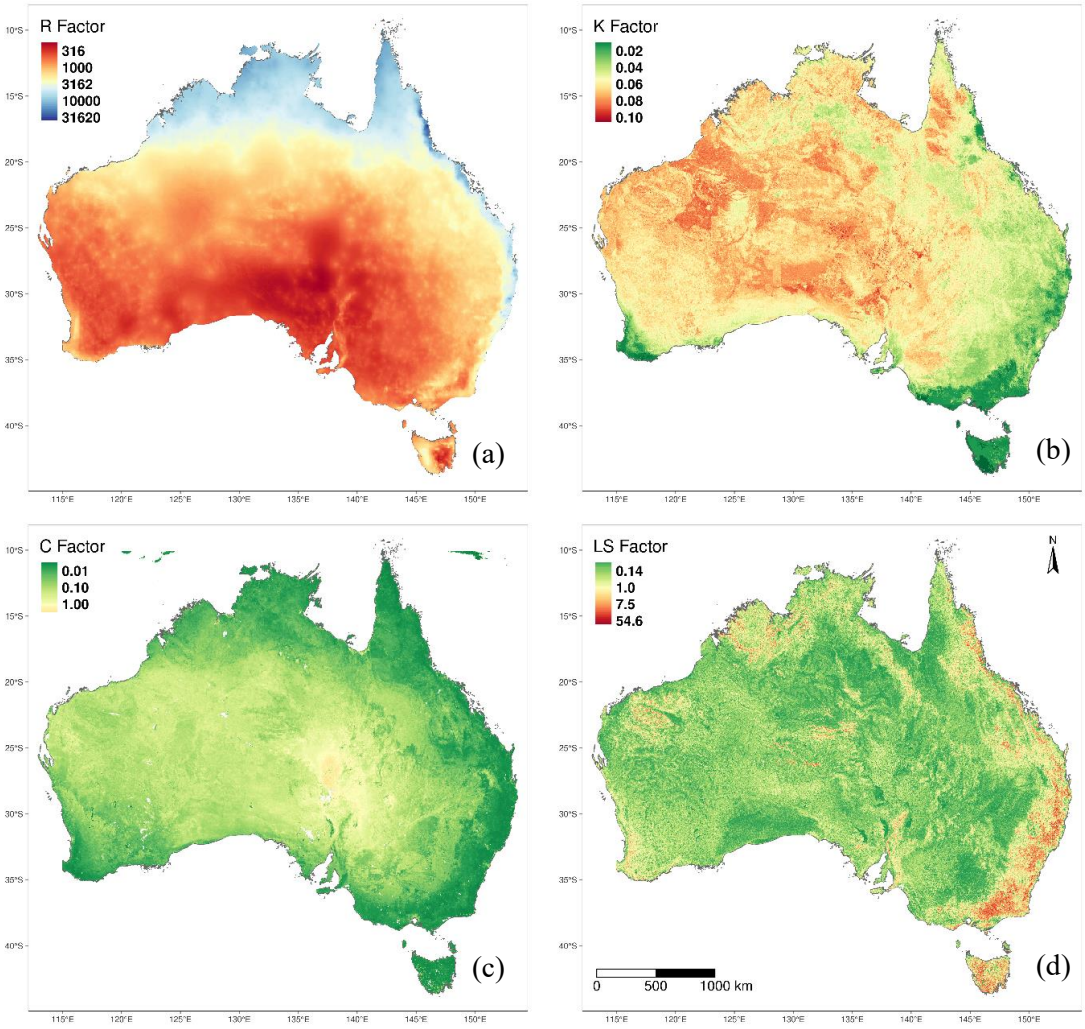


Figure 5-1 Maps of the RUSLE factors: Rainfall erosivity factor, Soil erodibility factor, Cover management factor, Slope length and steepness factor

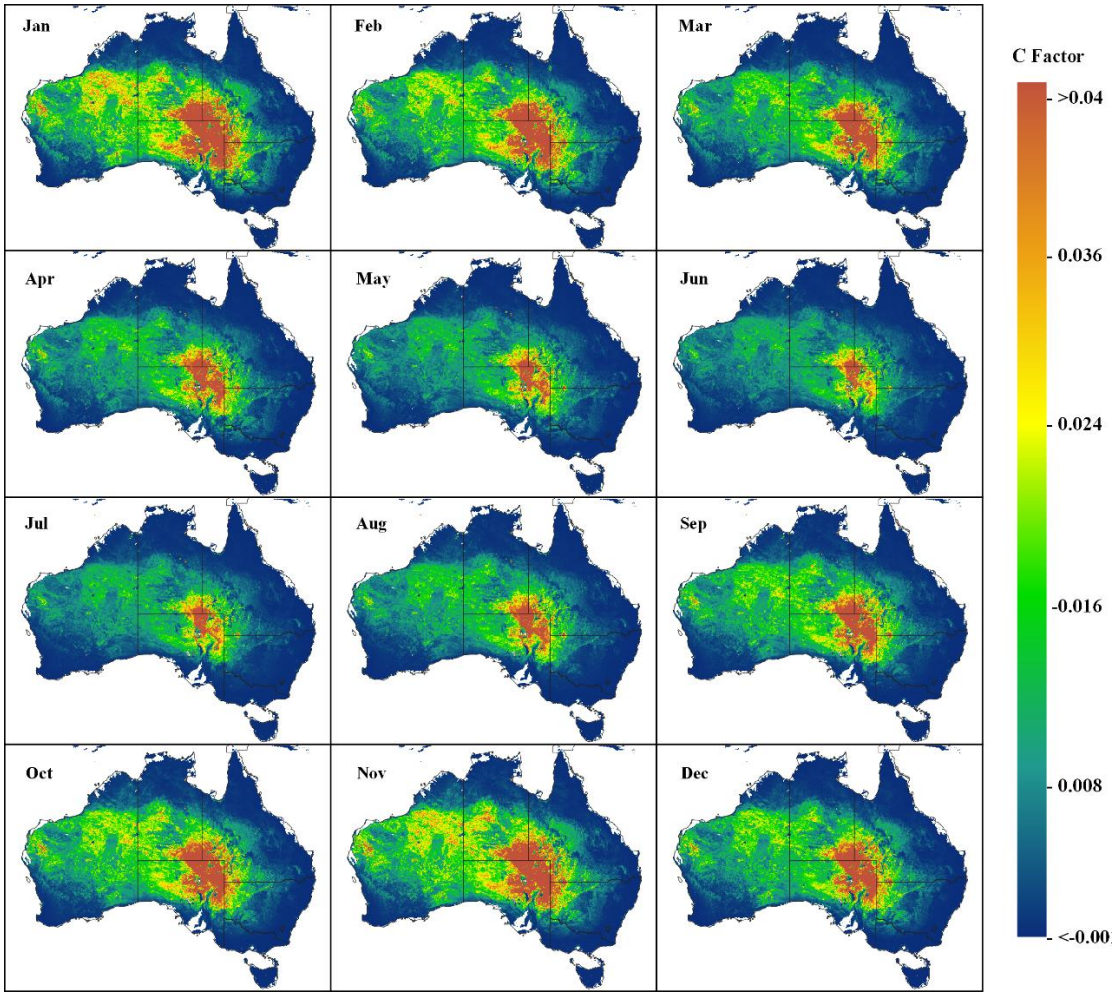


Figure 5-2 Monthly C-factor based on MODIS fractional vegetation cover in 2001-2020.

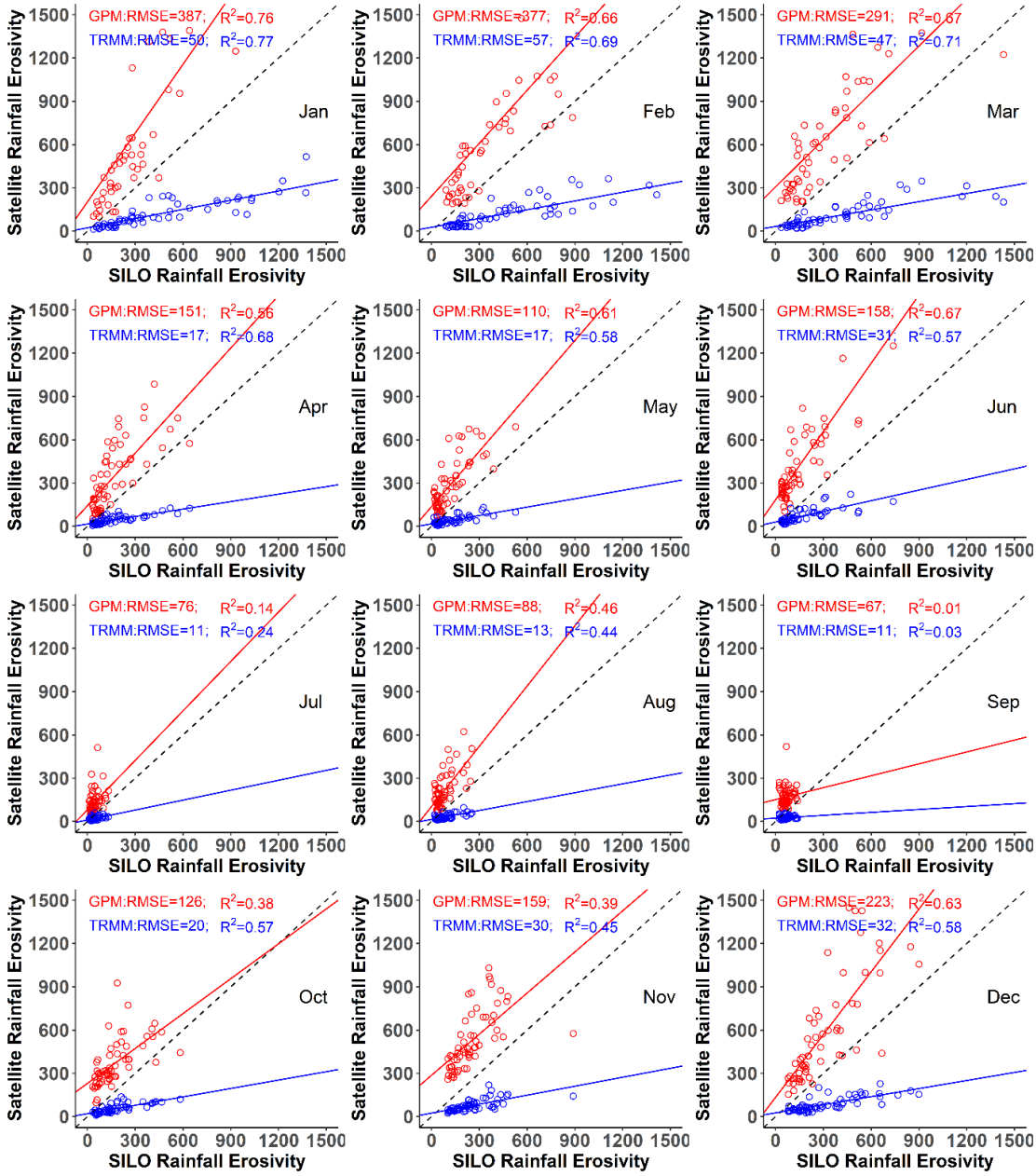


Figure 5-3 Comparison between the R factor values derived from SILO and GPM and TRMM for 12 months along the Great Dividing Range in south-eastern Australia

5.3.2 Assessment and comparison of two wind erosion model outputs with DustWatch

We compared the two models, RWEQ and Albedo, against each other and against the PM10 concentrations measured by DustWatch at Tibboburra from 2009 to 2020. Using the GLDAS dataset as model input, Figure 5-4 shows that the monthly wind erosion fluxes in

RWEQ model are in good agreement with that of Albedo-based model, while the monthly wind erosion estimates in albedo-based model are relatively lower than the model results of RWEQ. In one financial year (from July to next June) as an example, the wind erosion generally starts to rise in July, reaches to its highest value in January, and then declines until it reaches the lowest values in May.

The Australian Bureau of Meteorology (BoM) climate summaries were used to identify wet and dry years. 2010 to 2012 where wet years while 2001-2009 and 2017-2019 had hot and dry conditions in south-eastern Australia, of which 2009 and 2019 were two particularly dry and dusty years. For the two dry years, the measured fluxes are less than or equal to the model results from RWEQ in 2009 and 2019, respectively. By contrast, the measured fluxes are higher than the albedo-based model results. In other words, the measured fluxes should not be greater than the model results. It should be noted that the DWN flux measurements do not adequately represent the areal wind erosion, for example, fine dust may travel thousands of kilometres or remain in air until it is washed out by rainfall, the dust storm in the atmosphere measured by DustWatch are mainly derived from near-surface wind erosion. We found that there was a peak value (~246 mg/m2/month) in September 2009. Based on our daily wind erosion results, the extreme value is mainly coming from the wind erosion on Sep 23, 2009 (a dust storm named Red Dawn), with a maximum wind speed of 12.5 m/s. However, this anomaly was not well represented in the Albedo-based model. In fact, the spike erosion fluxes of 2009 in albedo-based model shows similar pattern of other dry years. While the RWEQ model can better capture the extreme dust storm in 2009. To further confirm this case, however, the RWEQ model results in 2019 does not show superior performance than the albedo-based model. This may be the reason that the observations at hot and dry year such as 2019 was affected by other factors like suspension from other places. In addition, from the beginning of 2010 to later 2011, the observed flux remained unchanged and always lower than the simulated results. As the La Nina in 2010 and 2011 breaks the drought, south-eastern Australia enters in wet year. The precipitation increased soil moisture and affects the threshold wind speed of model.

To evaluate the impacts of different data source on wind erosion results, we compared the model results derived from ERA5 datasets with measured wind erosion fluxes (Figure 5-5). We can see that the temporal variation pattern of two model results can match that of measured fluxes. Though the wind erosion fluxes of RWEQ model is equal to or smaller than that of the Albedo-based model, it clearly shows that RWEQ model can capture the extreme value in 2009. However, both the RWEQ and albedo-based model do not capture the measured fluxes in 2019, which may be influenced by the mega-fire of Australia followed by smoke and dust. Overall, using ERA5 datasets as model input, the simulated wind erosion fluxes are significantly less than that of GLDAS.

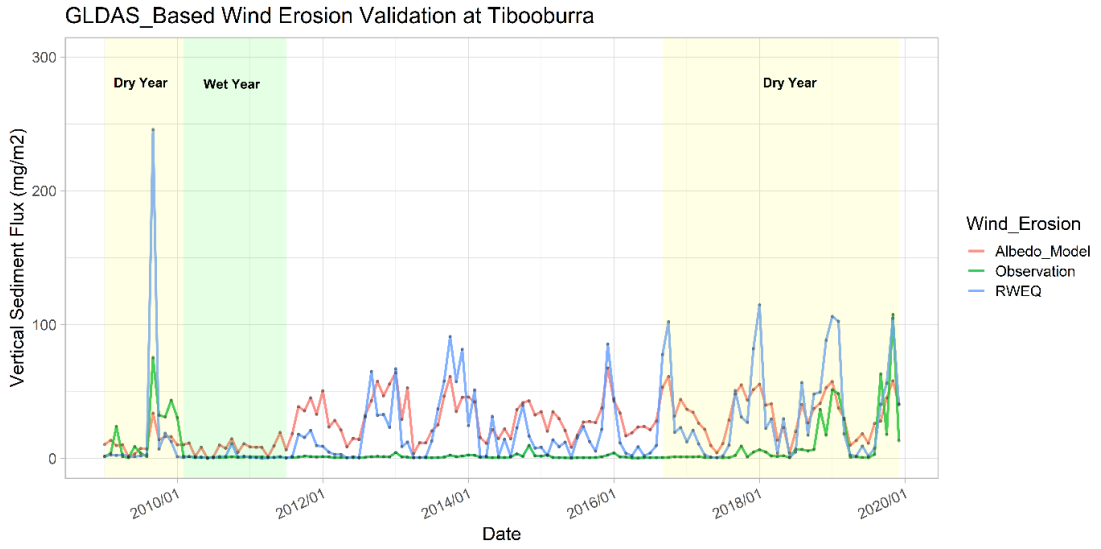


Figure 5-4 Monthly wind erosion values from the albedo-based model, the RWEQ model, and the observations at Tibooburra site are compared for the period 2009 to 2019 using the GLDAS dataset as model input

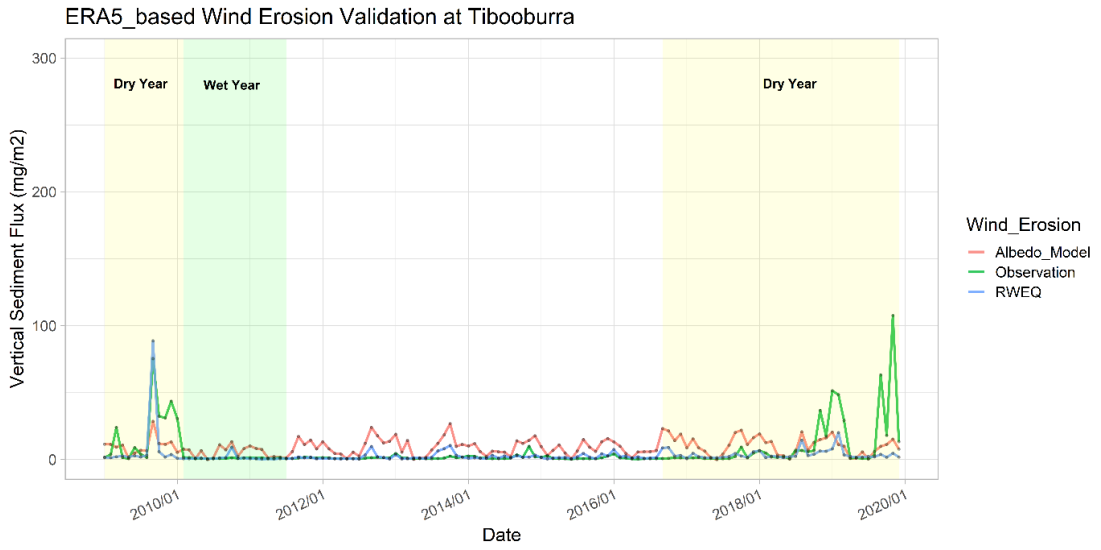


Figure 5-5 Monthly wind erosion values from the albedo-based model, the RWEQ model, and the observations at Tibooburra site are compared for the period 2009 to 2019 using the ERA5 dataset as model input

5.3.3 Monthly and annually wind-water erosion maps

We produced the water and wind erosion monthly time-series maps for the period 2001-2020 across Australia. Spatially, the water erosion process across the continent is mainly concentrated in the Great Dividing Range along eastern coastal Australia, Flinders Range in south Australia (SA), Lake Eyre in central Australia, Hamersley Range in western Australia (WA), Barkly tableland and Arnhem Land in North Territory (NT) and western coastal of Tasmania (TAS) (Figure 5-6). Moreover, the range of monthly water erosion varies by seasons across Australia. Over summer, the water erosivity is apparent across the most of Australia. The summer water erosivity is more evident in northern and southeastern coastal regions. While water erosivity in winter was relatively smaller because of reduced rainfall. The most obvious winter water erosivity is observed in Alpine region in Victoria and western coastal of Tasmania. Figure 5-7 further demonstrates that WA is the top region by soil loss of water erosion in Australia. The top 3 regions also include Queensland (QLD) and NT. Through the year, the highest month of water erosion is Jan and the lowest is August. Taking WA as an example, water erosion in summer counts for 74% of total annual erosion, water erosion in spring and

winter counts for 7% and 3% for each and winter counts for 16%. Change on the intra-annual trend of water erosion in other states is almost the same as WA, except for Tasmania where is highest in winter and lowest in summer.

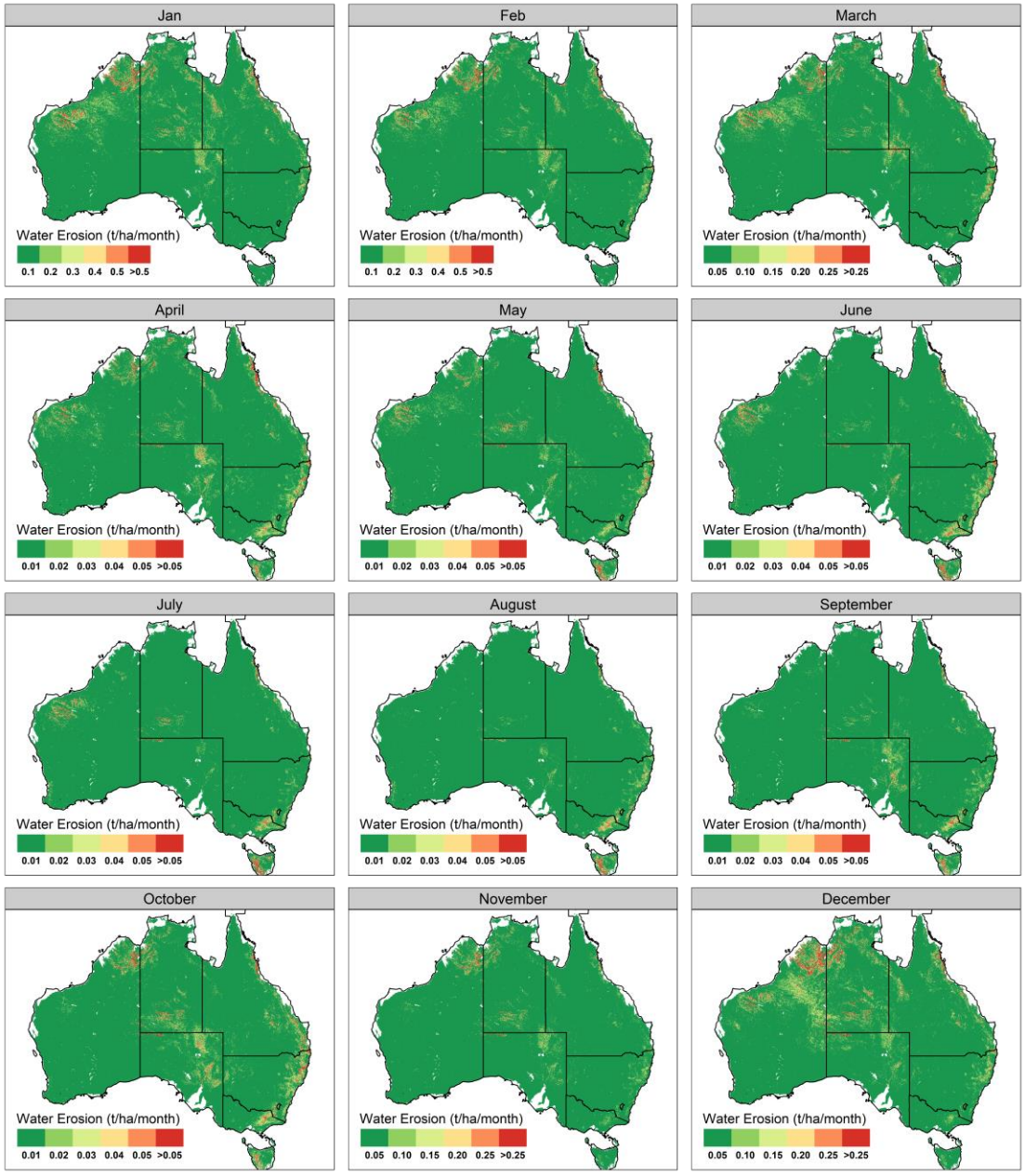


Figure 5-6 Monthly water erosion based on SILO in 2001-2020

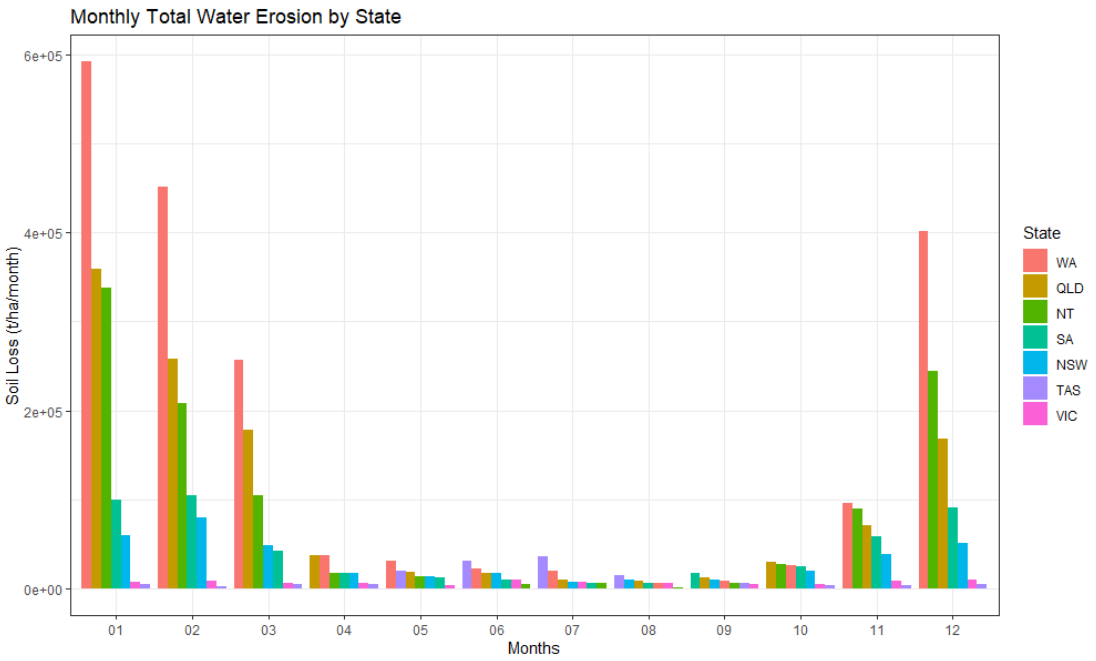


Figure 5-7 Monthly water erosion by State in 2001-2020

The albedo model outputs are shown in Figure 5-8. The highest region with wind erosion is Nullarbor Plain, which straddles SA and WA and lies directly north of the Great Australia Bight, followed by the Great Sandy Desert in WA, and the Lake Eyre basin. Figure 10 also shows that the strong monthly variations, wind erosion is typically the greatest in Austral summer months (December, January, February) and least in Austral winter months (June, July, August). Comparing the wind erosion amount at the state scale, the arid and semi-arid parts of WA and SA experienced higher levels of wind erosion than eastern states (seen in Figure 5-9). It is also worth noted that Eyre Peninsula (SA) in Figure 5-8 shows strong wind erosion activities compared to that of Figure 5-10.

The RWEQ outputs are shown in Figure 5-10 and shows the RWEQ model has over 10 times monthly wind erosion outputs than that of the Albedo-based model. As Chappell et al. (2018) explains, fractional cover does not capture the effects of cover roughness and can lead to overestimates of wind erosion. This is because fractional ignores the sheltering effects roughness provides to bare ground downwind of the roughness element. Figure 5-11 shows that the monthly wind erosion outputs from RWEQ model are not only relatively higher but have

great seasonal differences than that of Albedo-based model (seen in Figure 5-9). That difference may be due to the relatively stable albedo across various land types. For example, the highest albedo for desert is about 0.4, and the intermediate albedo for crops and grasslands is around 0.2, and the lowest for forests is about 0.1. While the fractional vegetation cover used in RWEQ model shows greater variation among different landscapes. Furthermore, we also find that wind erosion in SA is greater than WA from August to November (seen in Figure 5-11).

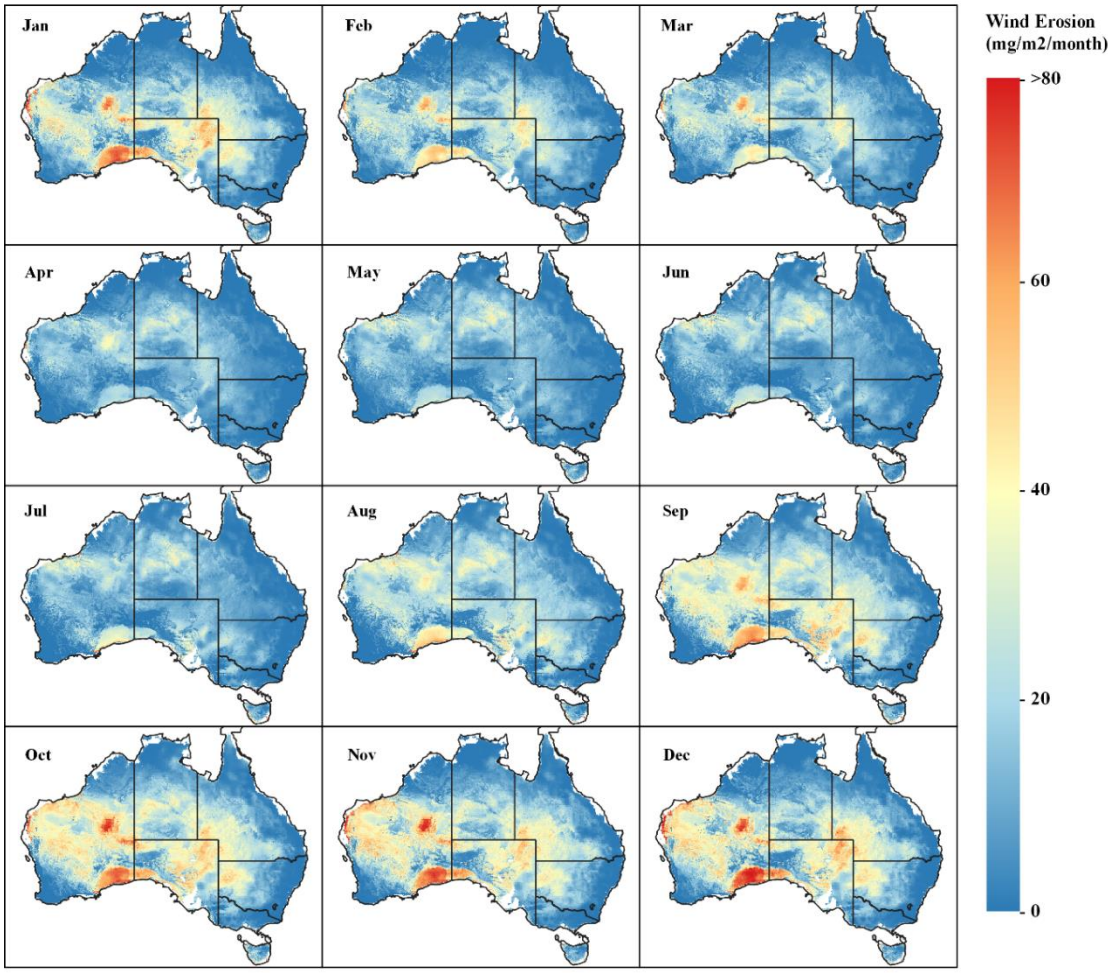


Figure 5-8 Monthly wind erosion based on Albedo-based model in 2001-2020

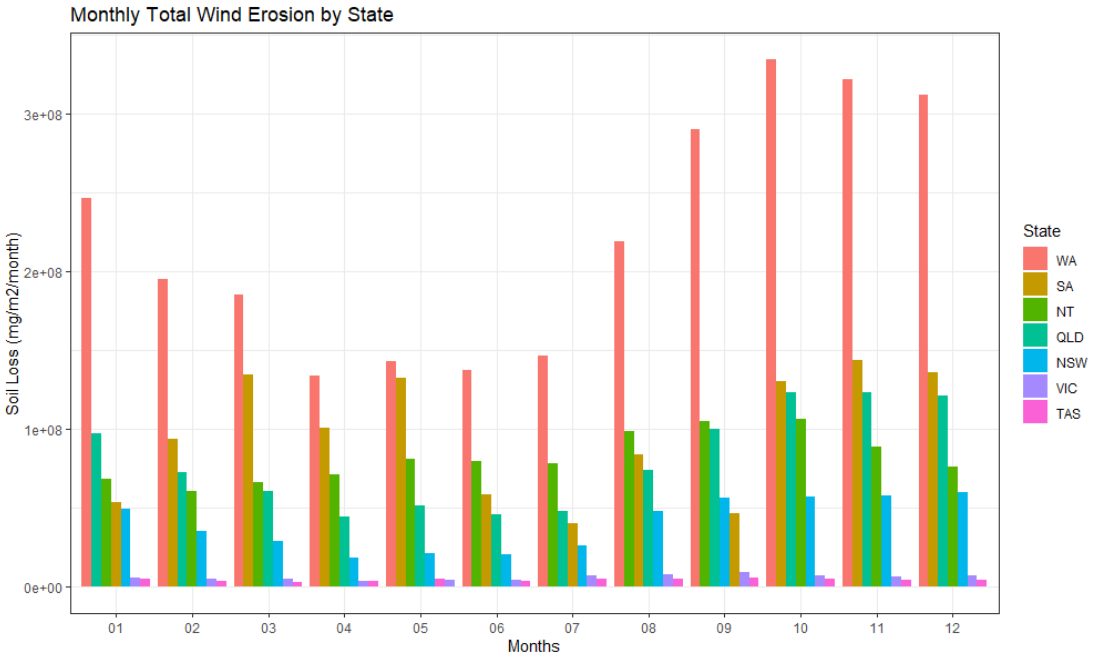


Figure 5-9 Monthly wind erosion by State based on Albedo-based model in 2001-2020

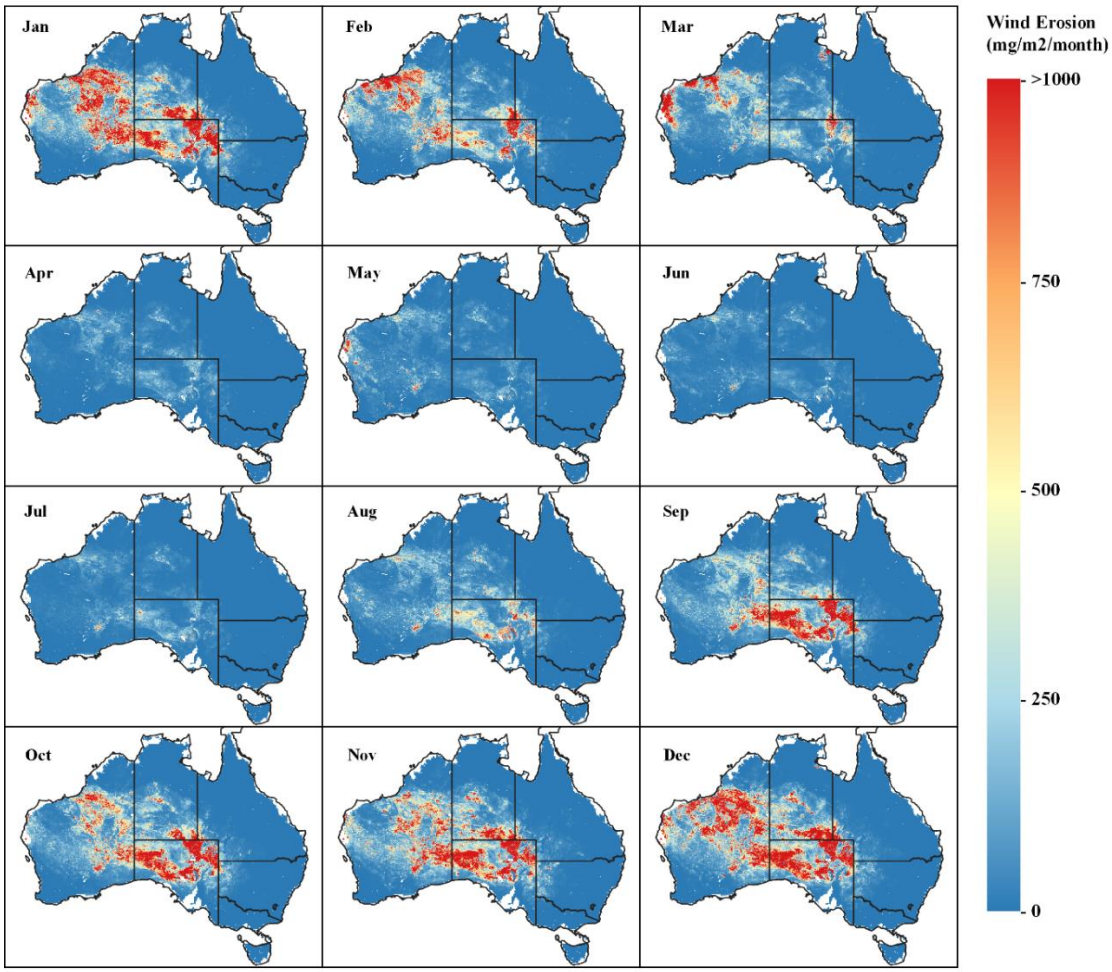


Figure 5-10 Monthly wind erosion based on RWEQ model in 2001-2020

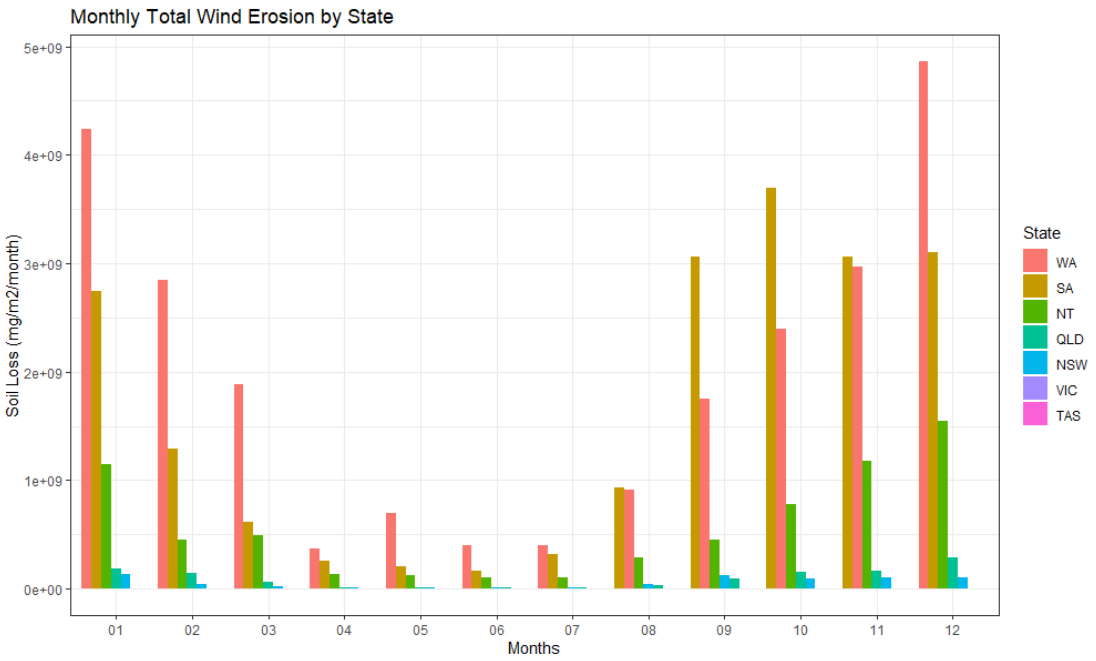


Figure 5-11 Monthly wind erosion by State based on RWEQ in 2001-2020

Maps of the RUSLE and RWEQ erosion and uncertainty maps both at 500 m pixel are shown in Figure 5-12. The uncertainty calculation method refers to Teng et al. (2016). We note that the largest water erosion rates existed around the Kimberley plateau in northern Australia, Canning basin in western Australia, Simpson Desert in central Australia, coastal wet tropics in western Australia and southeast slopes in New South Wales and Victoria. We also found that the rangeland areas of inland Australia experienced relatively larger water erosion loss, while the corresponding uncertainty of inland Australia are about four time larger than other regions, and these larger uncertainties should be come from the Rainfall erosivity estimates for dryland areas. Also, wind erosion loss was relatively larger in the rangeland areas where the vegetation cover is sparse. One explanation is that the soils in the rangeland areas are much easier to be

detached and transported by water and wind.

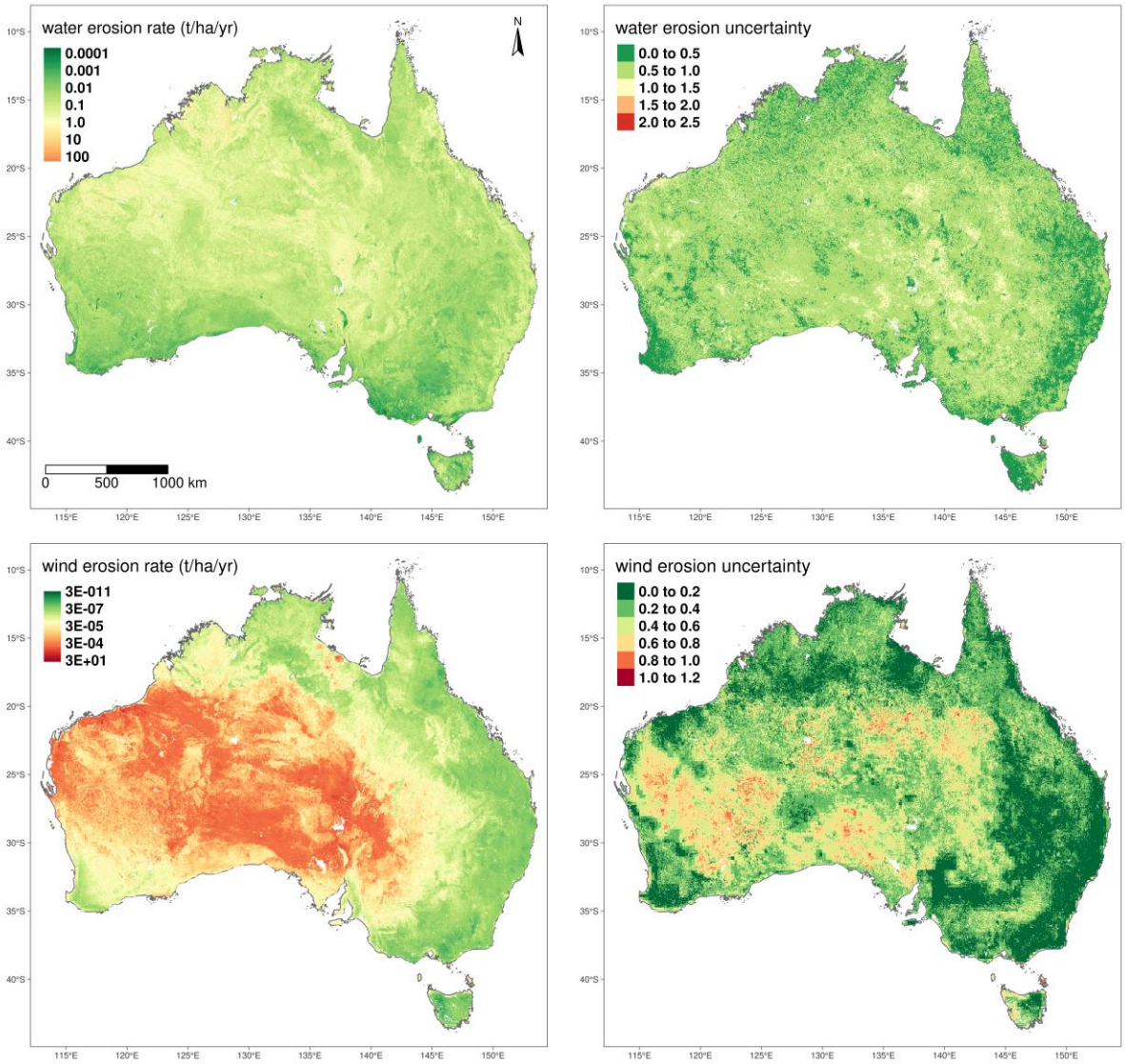


Figure 5-12 Annual water and wind erosion and uncertainty based on RUSLE and RWEQ in 2001-2020.

5.4 Discussion

5.4.1 Water and wind erosion explorer

In this study, we estimate the soil loss using the RUSLE and RWEQ model. RUSLE and RWEQ model are then translated into GEE environment to enable large-scale spatio-temporal soil loss simulations. Also, we build a web tool with user-friendly interface for public users.

This web tool provides an instant access to continental-scale remote sensing big data obtained for soil, vegetation, weather, topography, and other biophysical factors as input and strong computation capability of GEE. The application of water and wind models across Australia provides thorough datasets in states like WA and NT where very high soil loss rate cannot be neglected. Moreover, with this app link, the user can plot several figures displaying the time-series of soil loss records and four sub-factor maps (R, K, LS, C Factor in RSULE; SW, WF, K, MF, CF factor in RWEQ). The information can be exported (e.g., CSV files) for further analysis. In such a case, analysis-ready water and wind erosion application proves to be a practical and economical way for real-time and human-interactive visualization. Water and wind erosion explorer is freely available from the authors for educational and academic purposes at https://github.com/geogismx/WaterWindErosion. The water and wind erosion explorer is publicly available. While we have focused on the soil loss of Australia, users can also define their own research area and produce their erosion outcomes.

5.4.2 Underlying drivers for water and wind erosion changes

The reasons for water-winter erosion changes are complex and a result of combination of several factors including weather factors (the key driving force), vegetation cover (the key resistance), soil physical properties (determining soil erodibility or resistance to erosion), land cover types (influencing soil surface roughness, soil particles) and the development of wind profiles. In RUSLE model, rainfall erosivity is the most dominant agent, changes in rainfall can largely impact changes in water erosion trends (Du et al., 2015; Sun et al., 2014). Generally, Australian rainfall in the past two decades experienced substantial inter-annual variability with short wet periods (2010–2011, 2020) and long-term dry periods (e.g., 2002–2009 and 2012–2019). Liu et al. (2020) found that Australian rainfall changes had a trend of short and frequent rainfall events, which meant most rainfall tend to evaporation rather than increasing soil moisture. As shown in Figure 2, our water erosion study was largely relied on the prediction of rainfall erosivity which had limitations associated with the smaller and frequent rainfall events, characterized by decreased rainfall intensity and increased rainfall probability. The limitations

point to the importance of using advanced machine learning to present the rainfall erosivity under different rainfall events. Additionally, the dry conditions and decrease in soil moisture impact the wind erosion. However, the water erosion changes trend was not always in agreement with rainfall changes. The bushfire-related significant vegetation cover declines also contribute to water erosion increase. A very strong water erosion peak in 2020 in NSW Sydney Drinking Catchment shows that the increase in rainfall and decrease in vegetation cover were highly in line with the increase in water erosion (Yang et al., 2020). Unlike changes in rainfall, long-term variability and seasonal pattern in near-surface wind speed for the past two decades are still poor understood.

Comparing to RUSLE subfactors in CSIRO data portal, we find that the C factor shows significant differences. The differences are from the different vegetation cover datasets. Our estimates used dynamic fractional vegetation cover to estimate C factor, whereas the C factor values from Teng et al. (2016) are just based on static land cover classification. The largest C factor values from Teng et al. (2016) cover all the arid and desert areas in central and western Australia. Our C factor estimates only have the largest values in the Lake Eyre basin of central Australia. In addition, despite the under-and overestimation of R factor values by GPM and TRMM, the multi-year average RMSE of R factor values (1527 $MJ{\cdot}mm{\cdot}ha^{-1}{\cdot}h^{-1}{\cdot}y^{-1}$ in GPM, 236 $MJ{\cdot}mm{\cdot}ha^{-1}{\cdot}h^{-1}{\cdot}y^{-1}$ in TRMM) in our study is much better than the RMSE of TRMM-based R factor values (2726 $MJ{\cdot}mm{\cdot}ha^{-1}{\cdot}h^{-1}{\cdot}y^{-1}$) from the literature (Teng et al., 2016). Lu et al. (2003) derived the water erosion in Australia during 1980-2000, and the mean water erosion in Australia is 11.77 $t{\cdot}ha^{-1}{\cdot}y^{-1}$, and the total erosion is 314.35 $t{\cdot}ha^{-1}{\cdot}y^{-1}$. Estimates of mean water erosion and total soil loss from 2002 and 2012 produced by Teng et al. (2016) are 7.2 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ and 100.22 $t{\cdot}ha^{-1}{\cdot}y^{-1}$. The uncertainty of mean water erosion range between 6.8 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ and 7.48 $t{\cdot}ha^{-1}{\cdot}y^{-1}$, the total soil loss is between 96.45 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ and 104.28 $t{\cdot}ha^{-1}{\cdot}y^{-1}$. Our overall estimates of erosion and total soil loss in Australia during 2000 and 2020 is 0.19 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ and 6.4 $t{\cdot}ha^{-1}{\cdot}y^{-1}$, respectively. The corresponding uncertainty of mean water erosion range between 0.09 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ and 0.22 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ the total soil loss is between 3.2 $t{\cdot}ha^{-1}{\cdot}y^{-1}$ and 7.5 $t{\cdot}ha^{-1}{\cdot}y^{-1}$. Our

estimates of erosion are smaller both in annual mean erosion and total soil loss. Chappell et al. (2011) provided the net erosion distribution of Australia of 1950-1990, and our map shows the similar erosion distribution in central Australia, while other estimates cannot capture the erosion in that area.

5.4.3 Limitations and Model Uncertainties

RUSLE model used in this study has been supported and calibrated by field monitoring and regional-scale models. Meanwhile, this study selected the RWEQ to predict the wind erosion of Australia as Albedo-based model tends to flatten out extremes. Note that most existing wind erosion models have not been calibrated and validated under wind tunnel measurements that can model the relationships between wind erosion and environmental conditions. This study is the first attempt to validate the empirical and process-based model RWEQ against real-time field observations and an albedo-based wind erosion model. However, RWEQ model with original parameters has not been calibrated against the local data of site observations collected at a regional scale. Therefore, uncertainties will be introduced when it is applied for monthly estimation at a regional scale because the parameters in the RWEQ vary in spatio-temporal scale. For example, two dynamic sub-factors that accounting vegetation cover and soil moisture effects in RWEQ need to be calibrated locally. The local coefficients in vegetation cover effects on wind erosion in AUSLEM was determined by wind tunnel experimentation at wind speed 18 m/s (Webb et al., 2009). RWEQ model also has limitation in estimating the rate of soil loss in suspension. The performance of RWEQ needs to improve after calibration for local condition, otherwise, the RWEQ model may underestimates erosion (Youssef et al., 2012). Therefore, the emphasis of future studies should be placed on improving the accuracy and developing new models. In addition, the future increase in the amount of high-resolution remote sensing big data will lead to improved estimation for vegetation, soil, topography, and other biophysical factors which will in turn improve the estimation of soil erosion and assist in calibrating and validating water and wind erosion models. It is acknowledged that the climate and vegetation cover are two key factors contributing to soil

erosion changes. One improvement area is to use high-quality FVC, radar-based rainfall erosivity, soil moisture estimation. For example, we can derive the monthly C-factor at 20 m resolution from Sentinel 2 with gap filling using image blending techniques. As the weather radar data across Australia are available, we can explore the use of radar-derived rainfall erosivity with calibration to model water erosion. Soil moisture from Australian Water Availability Project (AWAP) can also be used to replace the soil moisture data from global-scale GLDAS dataset. Furthermore, soil loss can be estimated using the strengths of process-based RUSLE and RWEQ models, keeping insights of the physical mechanism of erosion formation while exploiting the predictive capacity of machine learning with remote sensing big EO data.

5.5 Conclusion

In this study, we used the best available big EO data across Australia during the past two decades to derive the water and wind erosion datasets and analysed the spatio-temporal variability of soil erosion. We estimated rainfall erosivity across the country using satellite-based rainfall data and compared with the ground-based SILO data showing consistent spatial patterns. We also found a persistent high agreement among the two wind erosion models, but an apparent underestimation of the wind erosion by the albedo-based model. It is notable that RWEQ model is more sensitive to extreme climate. In practice, wind erosion model results demonstrate that two reanalysis data exhibit distinct erosion change. At the monthly scale, water and wind erosion has strong inter-annual seasonality. In the past two decades, the trends of annual water erosion vary among different states. While the annual wind erosion has increasing trend since 2010. Nevertheless, we first produce the water and wind erosion datasets across Australia which offer a perspective for understanding soil erosion and the changes in relation to climate, land, and soil conditions.

Chapter 6. Final conclusions and future research

6.1 Final conclusions

This study mainly focuses on the reconstruction of climate data at high spatiotemporal resolution based on big EO platform. The outcomes of this study provide crucial climate and environment information in China and Australia. The research methods and software presented in this study can be further applied to global.

We developed an improved GIS-based solar radiation model (STMSR, the spatial and temporal mountainous solar radiation model) that allows for treatment of high spatial and temporal variations in albedo, surrounding terrain shading and cloud cover for monitoring daily solar radiation at large scale. By comparison with other well-known GIS-based solar radiation models such as Solar Analyst in ArcGIS and r.sun in GRASS, our STMSR model showed better performance. The resulting estimates of global, direct, and diffuse solar radiation were validated with high estimation accuracy against the measured solar radiation data from 10 observation stations across Loess Plateau. Compared with other high-resolution solar radiation datasets, the global solar radiation presented in this paper has higher accuracy of daily solar radiation estimates over the Loess Plateau than other methods, generating higher R^2 and RMSE. Our STMSR model also has the potential to be applied globally for distributed modelling applications across a variety of landscapes.

We built an online tool based on a MODIS LST “hybrid” methodology to generate continuous daily maximum and minimum land surface temperature datasets in locations without observations and to provide the required remotely sensed inputs to air temperature prediction models. Changes in received solar energy among mountains inevitably affect the earth’s energy budget. We integrate mountain solar radiation and diverse remotely sensed vegetation indices to provide reliable temperature products over the TP. By comparing the performance of different machine learning techniques, we found the RF model performed best in predicting Tmax, Tmin, and Tmean. We expect the methodology we have developed can be

potentially useful for improving temperature datasets in mountainous regions around the globe, and thereby also improving climatic, environmental, hydrological, and ecological models.

We have developed a heat wave toolbox that has the ability to estimate past, current and future changes in heat waves at a continental scale. It uses a well-known heat wave framework considering intensity, frequency, magnitude, duration, and areal extent to explore the spatio-temporal evolution of heat wave severity and coverage. This study is the first attempt to estimate heat wave events across Australia using high spatio-temporal climate datasets. With these heat wave aspects from multi-source data and different methods, we were able to investigate the effects of scales, data quality and definition. We find that ERA5 datasets are the best in characterizing the heat wave events. In exploring the role of different methods on the identification of heat waves, we find that heatwave characteristics based on the Excess Heat Factor index provide more details on heatwave changes.

With the past 100 years of heat wave datasets, the HWA average mean values were calculated and used to estimate non-stationary return levels and return periods. We find that extreme heat wave events have much higher probability due to the effects of climate change. The heat wave event in 2019 may be more frequent in the coming decades. For the climate by the end of century, using heat wave metrics derived from a multi-model ensemble mean, we predict HWA to increase significantly during the two future periods and a larger fraction of southern Australia is projected to experience more extreme heat wave events. Furthermore, the patterns of change for HWD are opposite to those for HWA; northern Australia shows significant increases and southern Australia experience a moderate increase. The methodology and the cloud computing-based toolbox (HWT) is useful for dynamic visualization, extraction, and processing of complex heat wave events, and applicable anywhere in the world.

We used the best available big EO data across Australia during the past two decades to derive the water and wind erosion datasets and analysed the spatio-temporal variability of soil erosion. We estimated rainfall erosivity across the country using satellite-based rainfall data and compared with the ground-based SILO data showing consistent spatial patterns. We also

found a persistent high agreement among the two wind erosion models, but an apparent underestimation of the wind erosion by the albedo-based model. It is notable that RWEQ model is more sensitive to extreme climate. In practice, wind erosion model results demonstrate that two reanalysis data exhibit distinct erosion change. At the monthly scale, water and wind erosion has strong inter-annual seasonality. In the past two decades, the trends of annual water erosion vary among different states. While the annual wind erosion has increasing trend since 2010. Nevertheless, we first produce the water and wind erosion datasets across Australia which offer a perspective for understanding soil erosion and the changes in relation to climate, land, and soil conditions.

6.2 Future research

The aims of future research are to: (1) Assessment and comparison of different cloud cover model for mountainous solar radiation forecast; (2) To integrate Deep Learning with physical approaches for soil-vegetation-climate modelling; (3) To back forecast and future forecast soil organic carbon in Australia with Big EO datasets; (4) To mine the global heat wave under CMIP6 and explore the relationship between heat wave and soil moisture and climate indices.

Reference

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A. and Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. Scientific Data, 5: 171091.

Aguilar, C., Herrero, J. and Polo, M.J., 2010. Topographic effects on solar radiation distribution in mountainous watersheds and their influence on reference evapotranspiration estimates at watershed scale. Hydrology and Earth System Sciences, 14(12): 2479-2494.

Alexander, L.V. and Perkins, S.E., 2013. On the Measurement of Heat Waves. Journal of Climate, 26(13): 4500-4517.

Alexander, L.V. et al., 2006. Global observed changes in daily climate extremes of temperature and precipitation. Journal of Geophysical Research, 111(D5).

Alsamamra, H., Ruiz-Arias, J.A., Pozo-Vázquez, D. and Tovar-Pescador, J., 2009. A comparative study of ordinary and residual kriging techniques for mapping global solar radiation over southern Spain. Agric. For. Meteorol., 149(8): 1343-1357.

Amani, M. et al., 2020. Google Earth Engine Cloud Computing Platform for Remote Sensing Big Data Applications: A Comprehensive Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13: 5326-5350.

Anderson, M.C. et al., 2008. A thermal-based remote sensing technique for routine mapping of land-surface carbon, water and energy fluxes from field to regional scales. Remote Sensing of Environment, 112(12): 4227-4241.

Angstrom, A., 1927. Solar and terrestrial radiation. Quart. J. Roy. Meteorol. Soc., 50(120).

Argüeso, D. et al., 2015. Heatwaves affecting NSW and the ACT: recent trends, future projections and associated impacts on human health, NSW Office of Environment and Heritage, Sydney, Australia.

Atiqul Islam, M., Yu, B. and Cartwright, N., 2020. Assessment and comparison of five satellite precipitation products in Australia. Journal of Hydrology.

Bonavita, M. et al., 2021. Machine Learning for Earth System Observation and Prediction. Bulletin of the American Meteorological Society, 102(4): E710-E716.

Borrelli, P., Lugato, E., Montanarella, L. and Panagos, P., 2016. A New Assessment of Soil Loss Due to Wind Erosion in European Agricultural Soils Using a Quantitative Spatially Distributed Modelling Approach. Land Degradation & Development, 28(1): 335-344.

Borrelli, P. et al., 2017. An assessment of the global impact of 21st century land use change on soil erosion. Nature Communication, 8(1): 2013.

Borrelli, P. et al., 2020. Land use and climate change impacts on global soil erosion by water (2015-2070). PNAS, 117(36): 21994-22001.

Breiman, L., 2001. Random Forest. Machine Learning, 45: 5-32.

Bristow, K.L. and Campbell, G.S., 1984. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agricultural and Forest Meteorology, 31(2): 159-166.

Brock, T.D., 1981. Calculating solar radiation for ecological studies. Ecological modelling, 14(1-2): 1-19.
Brown, L. and Foster, G., 1987. Storm erosivity using idealized intensity distributions. Transactions of the ASAE, 30(2): 379-0386.
Bui, E., Hancock, G., Chappell, A. and Gregory, L., 2010. Evaluation of tolerable erosion rates and time to critical topsoil loss in Australia, Canberra, Australia.
Cai, D., You, Q., Fraedrich, K. and Guan, Y., 2017. Spatiotemporal Temperature Variability over the Tibetan Plateau: Altitudinal Dependence Associated with the Global Warming Hiatus. Journal of Climate, 30(3): 969-984.
Chappell, A., Viscarra Rossel, R.A. and Loughran, R., 2011. Spatial uncertainty of137Cs-derived net (1950s–1990) soil redistribution for Australia. Journal of Geophysical Research, 116(F4).
Chappell, A. and Webb, N.P., 2016. Using albedo to reform wind erosion modelling, mapping and monitoring. Aeolian Research, 23: 63-78.
Chappell, A. et al., 2018. Improving ground cover monitoring for wind erosion assessment using MODIS BRDF parameters. Remote Sensing of Environment, 204: 756-768.
Chappell, A. et al., 2019. Minimising soil organic carbon erosion by wind is critical for land degradation neutrality. Environmental Science & Policy, 93: 43-52.
Chen, R. et al., 2019. A hybrid CNN-LSTM model for typhoon formation forecasting. GeoInformatica, 23(3): 375-396.
Cheng, L., AghaKouchak, A., Gilleland, E. and Katz, R.W., 2014. Non-stationary extreme value analysis in a changing climate. Climatic change, 127(2): 353-369.
Christidis, N., Jones, G.S. and Stott, P.A., 2014. Dramatically increasing chance of extremely hot summers since the 2003 European heatwave. Nature Climate Change, 5(1): 46-50.
Coles, S., Bawa, J., Trenner, L. and Dorazio, P., 2001. An introduction to statistical modeling of extreme values, 208. Springer.
Crosson, W.L., Al-Hamdan, M.Z., Hemmings, S.N.J. and Wade, G.M., 2012. A daily merged MODIS Aqua–Terra land surface temperature data set for the conterminous United States. Remote Sensing of Environment, 119: 315-324.
Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A. and Hess, K.T., 2007. Random forests for classification in ecology. Ecology, 88(11): 2783-2792.
Daly, C., Conklin, D.R. and Unsworth, M.H., 2009. Local atmospheric decoupling in complex topography alters climate change impacts. International Journal of Climatology, 30(12): 1857-1864.
Dimri, A.P., Bookhagen, B., Stoffel, M. and Yasunari, T., 2019. Himalayan Weather and Climate and their Impact on the Environment. Himalayan Weather and Climate and their Impact on the Environment. Springer Nature.
Du, H., Xue, X., Wang, T. and Deng, X., 2015. Assessment of wind-erosion risk in the watershed of the Ningxia-Inner Mongolia Reach of the Yellow River, northern China. Aeolian Research, 17: 193-204.

Duan, A., Wu, G., Liu, Y., Ma, Y. and Zhao, P., 2012. Weather and climate effects of the Tibetan Plateau. Advances in Atmospheric Sciences, 29(5): 978-992.
Dubayah, R. and Rich, P.M., 1995. Topographic solar radiation models for GIS. Int. J. Geogr. Inf. Sci., 9(4): 405-419.
Eilers, P.H., 2003. A perfect smoother. Analytical chemistry, 75(14): 3631-3636.
Faeth, P., 1994. Building the Case for Sustainable Agriculture: Policy Lessons from India, Chile, and Chile, and the Philippines. Environment: Science and Policy for Sustainable Development, 36(1): 16-39.
FAO, I., 2015. Status of the world's soil resources (SWSR)–main report, Food and agriculture organization of the United Nations and intergovernmental technical panel on soils, Rome, Italy 650.
Farr, T.G. et al., 2007. The Shuttle Radar Topography Mission. Rev. Geophys., 45(2).
Feron, S. et al., 2019. Observations and Projections of Heat Waves in South America. Scientific reports, 9(1): 8173.
Fischer, E.M. and Schär, C., 2010. Consistent geographical patterns of changes in high-impact European heatwaves. Nature Geoscience, 3(6): 398-403.
Fotheringham, A.S., Crespo, R. and Yao, J., 2015. Geographical and Temporal Weighted Regression (GTWR). Geogr. Anal., 47(4): 431-452.
Freitas, S., Catita, C., Redweik, P. and Brito, M.C., 2015. Modelling solar potential in the urban environment: State-of-the-art review. Renewable and Sustainable Energy Reviews, 41: 915-931.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics: 1189-1232.
Fryrear, D. et al., 2000. RWEQ: Improved wind erosion technology. The wind erosion prediction system and its use in conservation planning, 55(2): 183-189.
Fu, P. and Rich, P.M., 2002. A geometric solar radiation model with applications in agriculture and forestry. Computers and Electronics in Agriculture, 37(1-3): 25-35.
Fu, P. et al., 2019. A physical model-based method for retrieving urban land surface temperatures under cloudy conditions. Remote Sensing of Environment, 230: 111191.
Gasch, C.K. et al., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D+ T: The Cook Agronomy Farm data set. Spatial Statistics, 14: 70-90.
Gerber, F. et al., 2018. Predicting missing values in spatio-temporal remote sensing data. IEEE Transactions on Geoscience and Remote Sensing, 56(5): 2841-2853.
Gilleland, E. and Katz, R.W., 2016. extRemes2.0: An Extreme Value Analysis Package inR. Journal of Statistical Software, 72(8).
Gomes, V.C.F., Queiroz, G.R. and Ferreira, K.R., 2020. An Overview of Platforms for Big Earth Observation Data Management and Analysis. Remote Sensing, 12(8).
Gorelick, N. et al., 2017a. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202: 18-27.
Gorelick, N. et al., 2017b. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ., 202: 18-27.

Guerschman, J.P. and Hill, M.J., 2018. Calibration and validation of the Australian fractional cover product for MODIS collection 6. Remote sensing letters, 9(7): 696-705.
Guerschman, J.P. et al., 2015. Assessing the effects of site heterogeneity and soil properties when unmixing photosynthetic vegetation, non-photosynthetic vegetation and bare soil fractions from Landsat and MODIS data. 161: 12-26.
Guo, H., 2017. Big Earth data: A new frontier in Earth and information sciences. Big Earth Data, 1(1-2): 4-20.
Ham, Y.G., Kim, J.H. and Luo, J.J., 2019. Deep learning for multi-year ENSO forecasts. Nature, 573(7775): 568-572.
Hashimoto, H. et al., 2019. High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States. International Journal of Climatology, 39(6): 2964-2983.
He, J. et al., 2020. The first high-resolution meteorological forcing dataset for land process studies over China. Scientific Data, 7(1): 1-11.
He, T., Liang, S. and Song, D., 2014. Analysis of global land surface albedo climatology and spatial-temporal variation during 1981–2010 from multiple satellite products. Journal of Geophysical Research: Atmospheres, 119(17): 10,281-10,298.
Heffernan, J.E., Stephenson, A.G. and Gilleland, E., 2016. Ismev: an introduction to statistical modeling of extreme values. R package version pp. 41.
Hersbach, H. et al., 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146(730): 1999-2049.
Hirsch, R.M., Slack, J.R. and Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data. Water Resources Research.
Hobday, A.J. et al., 2016. A hierarchical approach to defining marine heatwaves. Progress in Oceanography, 141: 227-238.
Hofierka, J., Su´ ri, M., 2002. The solar radiation model for Open source GIS: implementation and applications. In Proceedings of the Open Source GIS-GRASS Users Conference: 19.
Hossain, M., Rekabdar, B., Louis, S.J. and Dascalu, S., 2015. Forecasting the weather of Nevada: A deep learning approach. Proceedings of the International Joint Conference on Neural Networks: 2-7.
Iqbal, M., 1983. An introduction to solar radiation. New York Academic Press Inc.
Islam, M.A., Yu, B. and Cartwright, N., 2020. Assessment and comparison of five satellite precipitation products in Australia. Journal of Hydrology, 590.
Iziomon, M.G., Mayer, H., 2001. Performance of solar radiation models—a case study. Agric. For. Meteorol., 110(1): 1-11.
Jarrah, M., Mayel, S., Tatarko, J., Funk, R. and Kuka, K., 2020. A review of wind erosion models: Data requirements, processes, and validity. Catena, 187.
Jiang, C., Zhang, H., Zhang, Z. and Wang, D., 2019. Model-based assessment soil loss by wind and water erosion in China's Loess Plateau: Dynamic change, conservation effectiveness, and strategies for sustainable restoration. Global and Planetary Change, 172: 396-413.

Jiang, S., Zheng, Y. and Solomatine, D., 2020. Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. Geophysical Research Letters, 47(13).
Jones, D.A., Wang, W. and Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Australian Meteorological and Oceanographic Journal, 58(4): 233.
Jones, E.L., Rendell, L., Pirotta, E. and Long, J.A., 2016. Novel application of a quantitative spatial comparison tool to species distribution data. Ecological Indicators, 70: 67-76.
Kalra, A. and Ahmad, S., 2009. Using oceanic-atmospheric oscillations for long lead time streamflow forecasting. Water Resources Research, 45(3).
Kang, S. et al., 2010. Review of climate and cryospheric change in the Tibetan Plateau. Environmental Research Letters, 5(1): 015101.
Karydas, C.G., Panagos, P. and Gitas, I.Z., 2014. A classification of water erosion models according to their geospatial characteristics. International Journal of Digital Earth, 7(3): 229-250.
Katz, R.W., 2010. Statistics of extremes in climate change. climate change, 100(1): 71-76.
Kilibarda, M. et al., 2014. Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. Journal of Geophysical Research: Atmospheres, 119(5): 2294-2313.
Kok, J.F. et al., 2014. An improved dust emission model – Part 1: Model description and comparison against measurements. Atmospheric Chemistry and Physics, 14(23): 13023-13041.
Kong, D., Zhang, Y., Gu, X., Wang, D.J.I.J.o.P. and Sensing, R., 2019. A robust method for reconstructing global MODIS EVI time series on the Google Earth Engine. ISPRS Journal of Photogrammetry and Remote Sensing, 155: 13-24.
Kuang, X. and Jiao, J.J., 2016. Review on climate change on the Tibetan plateau during the last half century. Journal of Geophysical Research: Atmospheres, 121(8): 3979-4007.
Laflen, J.M., Elliot, W., Flanagan, D., Meyer, C. and Nearing, M., 1997. WEPP-predicting water erosion using a process-based model. Journal of Soil Water Conservation, 52(2): 96-102.
Leihy, R.I., Duffy, G.A., Nortje, E. and Chown, S.L., 2018. High resolution temperature data for ecological research and management on the Southern Ocean Islands. Scientific Data, 5: 180177.
Lewis, S.C. and Karoly, D.J., 2013. Anthropogenic contributions to Australia's record summer temperatures of 2013. Geophysical Research Letters, 40(14): 3705-3709.
Leys, J., Strong, C., Heidenreich, S. and Koen, T., 2018. Where She Blows! A Ten Year Dust Climatology of Western New South Wales Australia. Geosciences, 8(7).
Leys, J.F., Heidenreich, S.K., Strong, C.L., McTainsh, G.H. and Quigley, S., 2011. PM10 concentrations and mass transport during “Red Dawn” – Sydney 23 September 2009. Aeolian Research, 3(3): 327-342.
Li, B., Chen, Y. and Shi, X., 2020a. Does elevation dependent warming exist in high mountain Asia? Environmental Research Letters, 15(2).
Li, X., 2020. Heat wave trends in Southeast Asia during 1979-2018: The impact of humidity. Science of The Total Environment, 721: 137664.

Li, X., Zhou, Y., Asrar, G.R. and Zhu, Z., 2018a. Creating a seamless 1 km resolution daily land surface temperature dataset for urban and surrounding areas in the conterminous United States. Remote Sensing of Environment, 206: 84-97.
Li, X., Zhou, Y., Asrar, G.R. and Zhu, Z., 2018b. Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. Remote Sensing of Environment, 215: 74-84.
Li, Y. et al., 2020b. Big Data and Cloud Computing, Manual of Digital Earth, pp. 325-355.
Liu, D.L. et al., 2020. Characterizing spatiotemporal rainfall changes in 1960–2019 for continental Australia. International Journal of Climatology, 41(S1): 2420-2444.
Liu, J.D., Pan, T., Chen, D.L., Zhou, X.J., Yu Q., Gerald, N. F., Liu.D.L., Zou, X.T., Hans, W.L., D, J., Wu, D.R., Shen, Y.B., 2017. An Improved Ångström-Type Model for Estimating Solar Radiation over the Tibetan Plateau. Energies, 10(7).
Liu, M., Bárdossy, A., Li, J. and Jiang, Y., 2012. GIS-based modelling of topography-induced solar radiation variability in complex terrain for data sparse region. Int. J. Geogr. Inf. Sci., 26(7): 1281-1308.
Liu, X. et al., 2009. Calibration of the Ångström–Prescott coefficients (a, b) under different time scales and their impacts in estimating global solar radiation in the Yellow River basin. Agric. For. Meteorol, 149(3-4): 697-710.
Louche, A., Notton, G., Poggi, P. SIMONNOT, G., 1991. Correlations for direct normal and global horizontal irradiation on a French Mediterranean site. Sol. Energy, 46(4): 5.
Lu, H. et al., 2003. Predicting sheetwash and rill erosion over the Australian continent. Australian journal of soil research, 41(6).
Lü, Y. et al., 2012. A Policy-Driven Large Scale Ecological Restoration: Quantifying Ecosystem Services Changes in the Loess Plateau of China. PLoS ONE, 7(2).
Luo, Q., 2011. Temperature thresholds and crop production: a review. Climatic Change, 109(3-4): 583-598.
Lyon, B., Barnston, A.G., Coffel, E. and Horton, R.M., 2019. Projected increase in the spatial extent of contiguous US summer heat waves and associated attributes. Environmental Research Letters, 14(11).
Ma, F., Yuan, X., Jiao, Y. and Ji, P., 2020. Unprecedented Europe Heat in June–July 2019: Risk in the Historical and Future Context. Geophysical Research Letters, 47(11).
Mallick, K. et al., 2014. A Surface Temperature Initiated Closure (STIC) for surface energy balance fluxes. Remote Sensing of Environment, 141: 243-261.
Manabe, S. and Terpstra, T., 1974. The effects of mountains on the general circulation of the atmosphere as identified by numerical experiments. Journal of the atmospheric Sciences, 31(1): 3-42.
McKenzie, N., Jacquier, D., Ashton, L. and Cresswell, H., 2000. Estimation of soil properties using the Atlas of Australian Soils.
McKenzie, N.J. et al., 2017. Priorities for improving soil condition across Australia's agricultural landscapes, CSIRO, Australia.

Mészároš, I. and Miklánek, P., 2006. Calculation of potential evapotranspiration based on solar radiation income modeling in mountainous areas. Biologia, 61(19).
Metz, M., Andreo, V. and Neteler, M., 2017. A New Fully Gap-Free Time Series of Land Surface Temperature from MODIS LST Data. Remote Sensing, 9(12).
Meyer, H. et al., 2016. Mapping daily air temperature for Antarctica based on MODIS LST. Remote Sensing, 8(9): 732.
Meyer, H., Reudenbach, C., Hengl, T., Katurji, M. and Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environmental Modelling & Software, 101: 1-9.
Middleton, N., 2019. Variability and Trends in Dust Storm Frequency on Decadal Timescales: Climatic Drivers and Human Impacts. Geosciences, 9(6).
Moreno-Martínez, Á. et al., 2018. A methodology to derive global maps of leaf traits using remote sensing and climate data. Remote sensing of environment, 218: 69-88.
Nairn, J.R. and Fawcett, R.J., 2015. The excess heat factor: a metric for heatwave intensity and its use in classifying heatwave severity. International journal of environmental research and public health, 12(1): 227-253.
Noi, P., Degener, J. and Kappas, M., 2017. Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data. Remote Sensing, 9(5).
Olaru, C., Wehenkel, L.J.F.s. and systems, 2003. A complete fuzzy decision tree technique. Fuzzy sets and systems, 138(2): 221-254.
Ouyang, X. et al., 2017. Validation and analysis of long-term AATSR land surface temperature product in the Heihe River basin, China. Remote Sensing, 9(2): 152.
Pan, Z., Hu, Y. and Cao, B., 2017. Construction of smooth daily remote sensing time series data: a higher spatiotemporal resolution perspective. Open Geospatial Data, Software and Standards, 2(1).
Pedamkar, P., 2020. Difference Between Big Data vs Data Science, https://www.educba.com/big-data-vs-data-science/.
Pepin, N. et al., 2015. Elevation-dependent warming in mountain regions of the world. Nature Climate Change, 5(5): 424-430.
Pepin, N. et al., 2019. An Examination of Temperature Trends at High Elevations Across the Tibetan Plateau: The Use of MODIS LST to Understand Patterns of Elevation-Dependent Warming. Journal of Geophysical Research: Atmospheres, 124(11): 5738-5756.
Perkins-Kirkpatrick, S.E. and Gibson, P.B., 2017. Changes in regional heatwave characteristics as a function of increasing global temperature. Scientific Reports, 7(1): 12256.
Perkins, S.E., 2015. A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at the global scale. Atmospheric Research, 164-165: 242-267.
Perkins, S.E., Alexander, L.V. and Nairn, J.R., 2012. Increasing frequency, intensity and duration of observed global heatwaves and warm spells. Geophysical Research Letters, 39(20).

Pi, H., Sharratt, B., Feng, G. and Lei, J., 2017. Evaluation of two empirical wind erosion models in arid and semi-arid regions of China and the USA. Environmental Modelling & Software, 91: 28-46.
Pintor, B.H. et al., 2015. Solar Energy Resource Assessment Using R. SUN In GRASS GIS And Site Suitability Analysis Using AHP For Groundmounted Solar Photovoltaic (PV) Farm In The Central Luzon Region (Region 3), Philippines, Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings, pp. 3.
Podestá, G.P., Núñez, L., Villanueva, C.A. and Skansi, M.a.A., 2004. Estimating daily solar radiation in the Argentine Pampas. Agric. For. Meteorol., 123(1-2): 41-53.
Purich, A. et al., 2014. More Frequent, Longer, and Hotter Heat Waves for Australia in the Twenty-First Century. Journal of Climate, 27(15): 5851-5871.
Qin, J. et al., 2015. An efficient physically based parameterization to derive surface solar irradiance based on satellite atmospheric products. J. Geophys. Res. D: Atmos., 120(10): 4975-4988.
Qin, J., Yang, K., Liang, S. and Guo, X., 2009. The altitudinal dependence of recent rapid warming over the Tibetan Plateau. Climatic Change, 97(1-2): 321.
Qiu, J., 2008. The third pole. Nature, 454(24): 393-396.
Quinlan, J.R., 1992. Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence: 343-348.
Raei, E., Nikoo, M.R., AghaKouchak, A., Mazdiyasni, O. and Sadegh, M., 2018. GHWR, a multi-method global heatwave and warm-spell record and toolbox. Scientific data, 5: 180206.
Rahmati, O., Tahmasebipour, N., Haghizadeh, A., Pourghasemi, H.R. and Feizizadeh, B., 2017. Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion. Geomorphology, 298: 118-137.
Rahmstorf, S. and Coumou, D., 2011. Increase of extreme events in a warming world. Proceedings of the National Academy of Sciences, 108(44): 17905-9.
Ramanathan, V. et al., 2007. Warming trends in Asia amplified by brown cloud solar absorption. Nature, 448(7153): 575-8.
Rashmi, K.V. and Gilad-Bachrach, R., 2015. DART: Dropouts meet Multiple Additive Regression Trees, AISTATS, pp. 489-497.
Ren, X. et al., 2021. Deep Learning-Based Weather Prediction: A Survey. Big Data Research, 23.
Ribatet, M., Singleton, R. and Team, R.C.J.R.p.v., 2011. SpatialExtremes: modelling spatial extremes. 1.8-1.
Rodell, M. et al., 2004a. The global land data assimilation system. Bull. Amer. Meteor. Soc., 85(3): 381-394.
Rodell, M. et al., 2004b. The global land data assimilation system. Bulletin of the American Meteorological Society, 85(3): 381-394.
Romano, F. et al., 2018. Improvement in Surface Solar Irradiance Estimation Using HRV/MSG Data. Remote Sensing, 10(8).

Roupioz, L., Jia, L., Nerry, F. and Menenti, M., 2016. Estimation of Daily Solar Radiation Budget at Kilometer Resolution over the Tibetan Plateau by Integrating MODIS Data Products and a DEM. Remote Sensing, 8(6): 504.
Ruiz-Arias, J.A., Tovar-Pescador, J., Pozo-Vázquez, D. and Alsamamra, H., 2009. A comparative analysis of DEM-based models to estimate the solar radiation in mountainous terrain. International Journal of Geographical Information Science, 23(8): 1049-1076.
Schaaf, C., Z. Wang., 2015. MCD43A3: MODIS/Terra and Aqua Albedo Daily L3 Global 500 m SIN Grid V006. NASA EOSDIS Land Processes DAAC.
Schär, C. et al., 2004. The role of increasing temperature variability in European summer heatwaves. Nature, 427(6972): 332-336.
Schlegel, R.W. and Smit, A.J., 2017. heatwaveR: A central algorithm for the detection of heatwaves and cold-spells. Journal of Open Source Software, 3(27): 1821.
Shan, L., Yang, X. and Zhu, Q., 2019. Effects of DEM resolutions on LS and hillslope erosion estimation in a burnt landscape. Soil Research, 57(7).
Shen, L., Mickley, L.J. and Gilleland, E.J.G.r.l., 2016. Impact of increasing heat waves on US ozone episodes in the 2050s: Results from a multimodel analysis using extreme value theory. Geophysical research letters, 43(8): 4017-4025.
Shi, X. et al., 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Advances in Neural Information Processing Systems: 802-810.
Stisen, S., Sandholt, I., Nørgaard, A., Fensholt, R. and Eklundh, L.J.R.S.o.E., 2007. Estimation of diurnal air temperature using MSG SEVIRI data in West Africa. Remote Sensing of Environment, 110(2): 262-274.
Stocker, T., (Ed.), 2014. Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
Sudmanns, M. et al., 2019. Big Earth data: disruptive changes in Earth observation data management and analysis? Int J Digit Earth, 13(7): 832-850.
Sun, Q., Miao, C., Duan, Q. and Wang, Y., 2015. Temperature and precipitation changes over the Loess Plateau between 1961 and 2011, based on high-density gauge observations. Global Planet. Change, 132: 1-10.
Sun, W., Shao, Q., Liu, J. and Zhai, J., 2014. Assessing the effects of land use and topography on soil erosion on the Loess Plateau in China. Catena, 121: 151-163.
Tabik, S., Villegas, A., Zapata, E.L. and Romero, L.F., 2012. A Fast GIS-tool to Compute the Maximum Solar Energy on Very Large Terrains. Procedia Comput. Sci., 9: 364-372.
Tamiminia, H. et al., 2020. Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. ISPRS Journal of Photogrammetry and Remote Sensing, 164: 152-170.
Tanarhte, M., Hadjinicolaou, P. and Lelieveld, J.J.C.R., 2015. Heat wave characteristics in the eastern Mediterranean and Middle East using extreme value theory. Climate Research, 63(2): 99-113.
Tang, W. et al., 2016. Retrieving high-resolution surface solar radiation with cloud parameters derived by combining MODIS and MTSAT data. Atmos. Chem. Phys., 16(4): 2543-2557.

Tatarko, J., Wagner, L. and Fox, F., 2019. The wind erosion prediction system and its use in conservation planning. The wind erosion prediction system and its use in conservation planning, 8: 71-101.
Telles, T.S., Dechen, S.C.F., Souza, L.G.A.d. and Guimarães, M.d.F., 2013. Valuation and assessment of soil erosion costs. Scientia Agricola, 70(3): 209-216.
Teng, H. et al., 2016. Assimilating satellite imagery and visible–near infrared spectroscopy to model and map soil loss by water erosion in Australia. Environmental Modelling & Software, 77: 156-167.
Todd R, L. and Dean L, U., 2003. Spatial estimation of air temperature differences for landscape-scale studies in montane environments. Agricultural and Forest Meteorology, 114(3-4): 141-151.
Trnka, M., Žalud, Z., Eitzinger, J. and Dubrovský, M., 2005. Global solar radiation in Central European lowlands estimated by various empirical formulae. Agricultural and Forest Meteorology, 131(1-2): 54-76.
Venter, Z.S., Brousse, O., Esau, I. and Meier, F., 2020. Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data. Remote Sensing of Environment, 242: 111791.
Viscarra Rossel, R.A. et al., 2015. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. Soil Research, 53(8).
Vogel, M.M., Zscheischler, J., Fischer, E.M. and Seneviratne, S., 2020. Development of future heatwaves for different hazard thresholds. Journal of Geophysical Research: Atmospheres, 125(9).
Wang, L., Qiu, X., Wang, P., Wang, X. and Liu, A., 2014a. Influence of complex topography on global solar radiation in the Yangtze River Basin. Journal of Geographical Sciences, 24(6): 980-992.
Wang, X. et al., 2014b. The dramatic climate warming in the Qaidam Basin, northeastern Tibetan Plateau, during 1961-2010. International Journal of Climatology, 34(5): 1524-1537.
Webb, N.P., McGowan, H.A., Phinn, S.R., Leys, J.F. and McTainsh, G.H., 2009. A model to predict land susceptibility to wind erosion in western Queensland, Australia. Environmental Modelling & Software, 24(2): 214-227.
Wiederholt, R. et al., 2019. A multi-indicator spatial similarity approach for evaluating ecological restoration scenarios. Landscape Ecology, 34(11): 2557-2574.
Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis: Principles and Applications. John Wiley and Sons Ltd.
Yang, K. et al., 2014. Recent climate changes over the Tibetan Plateau and their impacts on energy and water cycle: A review. Global and Planetary Change, 112: 79-91.
Yang, X., 2014. Deriving RUSLE cover factor from time-series fractional vegetation cover for hillslope erosion modelling in New South Wales. Soil Research, 52(3).
Yang, X., 2015. Digital mapping of RUSLE slope length and steepness factor across New South Wales, Australia. Soil Research, 53(2).

Yang, X., 2020. State and trends of hillslope erosion across New South Wales, Australia. Catena, 186.
Yang, X. et al., 2018. Digital mapping of soil erodibility for water erosion in New South Wales, Australia. 56(2): 158-170.
Yang, X. and Yu, B., 2015. Modelling and mapping rainfall erosivity in New South Wales, Australia. Soil Research, 53(2).
Yang, X. et al., 2020. Rapid Assessment of Hillslope Erosion Risk after the 2019–2020 Wildfires and Storm Events in Sydney Drinking Water Catchment. Remote Sensing, 12(22).
Yang, Y., Cai, W. and Yang, J., 2017. Evaluation of MODIS Land Surface Temperature Data to Estimate Near-Surface Air Temperature in Northeast China. Remote Sensing, 9(5).
Yeom, J.-M., Seo, Y.-K., Kim, D.-S. and Han, K.-S., 2016. Solar Radiation Received by Slopes Using COMS Imagery, a Physically Based Radiation Model, and GLOBE. Journal of Sensors, 2016: 1-15.
Yoo, C., Im, J., Park, S. and Quackenbush, L.J., 2018. Estimation of daily maximum and minimum air temperatures in urban landscapes using MODIS time series satellite data. ISPRS Journal of Photogrammetry and Remote Sensing, 137: 149-162.
You, Q., Min, J. and Kang, S., 2016. Rapid warming in the Tibetan Plateau from observations and CMIP5 models in recent decades. International Journal of Climatology, 36: 2660-2670.
Youssef, F., Visser, S., Karssenberg, D., Bruggeman, A. and Erpul, G., 2012. Calibration of RWEQ in a patchy landscape; a first step towards a regional scale wind erosion model. Aeolian Research, 3(4): 467-476.
Yu, B. and Rose, C., 1999. Application of a physically based soil erosion model, GUEST, in the absence of data on runoff rates I. Theory and methodology. Soil Research, 37(1): 1-12.
Zeng, Y., Qiu, X., He, Y. and Liu, C., 2008. Distributed modeling of diffuse solar radiation over rugged terrain of the yellow river basin. Chines J. Geophys-CH, 51(4).
Zeng, Y., Qiu, X., Liu, C. and Jiang, A., 2005. Distributed modelling of direct solar radiation of rugged terrain over the yellow river basin. Chines J. Geophys-CH, 60(4).
Zhang, G. et al., 2019. Impact of near-surface wind speed variability on wind erosion in the eastern agro-pastoral transitional zone of Northern China, 1982–2016. Agricultural and Forest Meteorology, 271: 102-115.
Zhang, H., Zhang, F., Ye, M., Che, T. and Zhang, G., 2016. Estimating daily air temperatures over the Tibetan Plateau by dynamically integrating MODIS LST data. Journal of Geophysical Research: Atmospheres, 121(19): 11,425-11,441.
Zhang, M. et al., 2020. Incorporating dynamic factors for improving a GIS-based solar radiation model. Transactions in GIS: 1-19.
Zhang, S., Yang, D., Yang, Y., Lei, H. and Fu, B., 2018. Excessive Afforestation and Soil Drying on China's Loess Plateau. J. Geophys. Res. G: Biogeosci., 123: 923–935.
Zhang, X. and Yang, F., 2004. RClimDex (1.0). Climate Research Branch.
Zhang, Y., Li, X. and Bai, Y., 2015. An integrated approach to estimate shortwave solar radiation on clear-sky days in rugged terrain using MODIS atmospheric products. Solar Energy, 113: 347-357.

Zhang, Z. and Li, J., 2020. Big climate data, Big Data Mining for Climate Change, pp. 1-18.
Zhao, G., Mu, X., Wen, Z., Wang, F. and Gao, P., 2013. Soil erosion, conservation, and eco-environment changes in the loess plateau of China. Land Degradation & Development, 24(5): 499-510.
Zhu, W., Lű, A. and Jia, S., 2013. Estimation of daily maximum and minimum air temperature using MODIS land surface temperature products. Remote Sensing of Environment, 130: 62-73.
Zhu, X., Zhang, Q., Xu, C., Sun, P. and Hu, P., 2019. Reconstruction of high spatial resolution surface air temperature data across China: A new geo-intelligent multisource data-based machine learning technique. Science Total of Environment, 665: 300-313.