

The role of climate datasets in understanding climate extremes.

Eresanya Emmanuel Olaoluwa^{1, 2, 3}, Israel R. Orimoloye^{4,5}, Olufemi Sunday Durowoju⁶, Mojolaoluwa Daramola^{6,8} Jesufemi Peter Olowe⁷, Folorunsho Akinleye⁷, Olasunkanmi Olorunsaye²

¹South China Sea Institute of Oceanology, Chinese Academy of Sciences, 164 West XingangRoad, Guangzhou 510301, People's Republic of China.

² Department of Marine Science and Technology, Federal University of Technology, P.M.B 704, Akure, Nigeria

³ Organization of African Academic Doctors (O_AAD)

⁴ Centre for Environmental Management, Faculty of Natural and Agricultural Sciences, University of the Free State, 339, Bloemfontein 9300, Republic of South Africa

⁵Risk & Vulnerability Science Centre, University of Fort Hare, Alice Campus, South Africa

⁶Department of Geography & Resource Management, Faculty of Social Sciences Osun State University, Okokum, Osun State, Nigeria

⁷ Department of Meteorology and Climate science, Federal University of Technology ,Akure, Nigeria

⁸ Institute of Geographic Sciences and Natural Resources Chinese Academy of Sciences.

Email: eresanyaemmanuel44@gmail.com, orimoloyeisrael@gmail.com
olufemi.durowoju@uniosun.edu.ng, femipeter46@gmail.com, folorunshoakinleye@gmail.com,
jollizle@gmail.com, olorunsayeolasunkanmi@gmail.com,

Abstract: The impact of climate extremes on the society has been of great concern to environmental scientist and policy makers. The destructive consequence attributed to natural hazards associated with climates extremes has been estimated to billions of dollars across the globe. To carry out a robust and effective researches that help to minimize or prevent the loss, detailed datasets of the past, present and future are needed. This will help to give an accurate prediction and early warning which is necessary for the policy making.

Keywords: Climate, Datasets, Extremes.

1. History of datasets, types and climates proxies

The climate of the earth is in a continual state of variation, changing from days to billions of years. These changes occur due to internal and external factors. To have an overview of the earth's climate history, a proper understanding of the present and past climate is necessary.

The usage of devices to measure climate parameters has only been in existence for about two centuries, therefore to estimate past climate, climate researchers make use of indirect methods of measurements known as climate proxies, which serves as natural archives of climate information (Jones and Mann, 2004). These proxies kept records of atmospheric configuration

and properties in form of written historical documentation or pictures that gives detailed description of the climate in their time (IPCC, 2007). Each of these climate proxies serves as an indirect methods of inferring past climate, which would require careful calibration and validation against instrument records of the same time.

Many paleoclimate studies make use of multiple proxy data in order to obtain comprehensive estimates (IPCC, 2007). Paleo climatological data is the first set of climate datasets that provided information about past climate (Osama *et al*, 2020). Paleo climatological data was first explored in the 20th century. The idea of the paleo climate was birthed in 1970 during the study of ice age and the possibility of its recurrence in the future (IPCC, 2007). The subject of paleoclimatology became important in order to understand the evolution of earth's climate from past to present (Jones and Mann, 2004, Osama *et al*, 2020). In the year 1990, when the first IPCC assessment was done, very little knowledge about climate variation prior to instrumental records was available, however about two decades later understanding improved greatly (IPCC, 2007). Paleo climatological datasets has helped to understand how climate has responded to climate forcing in the past and how it may respond to similar climate forcing in the future.

1.1 Historical Documents and Records

Another type of proxy data is the historical documents and records, which contains information about past climatic conditions. The climate information in this source are usually found in record written by Mariners, Farmers, traveller's diaries and Newspaper account of past weather. This records are usually in the form of occurrence date of drought, famine, frost, freezing of water bodies, snow and sea ice cover duration. Others may include phenological evidence such as the start and end of planting season, time of plant flowering, or date of harvest, all of which gives details of past climatic condition. Past and present graphical image of mountain glaciers has been used as evidence to support glacial retreat; which is a direct consequence of climate change. Despite the numerous uses of historical documentation, there are several limitations to its usage as a paleo-climate proxy. Historical documentations are mostly found within regions that have early writing traditions, therefore cannot be used to determine global climate conditions. Furthermore, the documentation of the climate or weather event is subject to the observer, as most of the writers would record the events in their local language, their timing (calendar) and based on their own perception without following any standard meteorological instrument's usage. Historical documentation mostly takes note of extreme climate events,

which may be considered as climate anomaly if careful consideration is not undertaken thereby providing false interpretation of events (Jones and Mann, 2004).

TREE RINGS RECORD

Records about past climatic conditions can be obtained from trees ring's archives containing long range of data, dating up to thousands of years (millennials); trees generally respond to changing climate by altering their growth pattern, which can be seen from the thickness of their rings. It is worthy to note that all trees can be used as proxies for climate studies, commonly used trees are found in the subpolar and multitude regions, these are majorly extratropical species, which can be cross-dated and chronologically developed. Several literatures consider tree ring as the most accurate source of paleo climate data that provide annual and seasonal accounts of climate condition (Jones and Mann, 2004, IPCC, 2007). Tree rings studies (dendroclimatology) can be used to produce climate information regarding temperature, precipitation, hydrology and fire.

ICE CORES RECORD

Thousands of years data of the past climatic conditions can also be obtained from ice cores of mountain glaciers, the polar ice caps or ice sheets (Greenland and Antarctica). Ice cores are found in the polar regions of Northern hemisphere and in the tropics and the sub tropics. Though found in small fraction over the earth's surface, it serve as complementary data sources to tree rings and corals. Ice cores are used as a source of paleoclimate data because falling snow traps air within tiny bubbles, which are latter compressed and converted to glacial ice with the air bubble still trapped. This ice core air bubbles contain information about the atmospheric composition of the atmosphere (rate of precipitation, fraction of melting ice and concentration of chemical constituents etc.). Seasonal pause in the accumulation of ice are used to establish chronology as observed layers or depths are connected with specific time space. Variation in temperature and precipitation can also be inferred from changes in the layering thickness.

CORALS RECORDS

Corals reefs found in the ocean can also provide information about past climate events, as they have been for millions of years and are exceedingly sensitive to changing climatic conditions. Corals are sensitive to temperature changes in the ocean, which may lead to bleaching, changes in pH levels, water pollution and runoff. Corals form their skeletons from calcium carbonate

in ocean water as they grow; the density and geochemical characteristic of the choral skeleton vary with the ocean properties (temperature, pH changes, fresh water influx, light, season, wave action and nutrient condition). This variation forms growth rings similar to tree growth rings and serves as the primary source of past annual and seasonal climate reconstruction from corals.

By observing trace elements, stable isotopes (oxygen isotopes) or variation in the density of the coral structure, scientist can establish a record of the sea-surface temperature, sea surface salinity and rainfall as well as the changes they have undergone overtime. Times of extreme events, environmental stress, disease outbreak and bleaching can also be identified, which is also helpful in determining condition that are harmful to the reef. Corals are found in both tropical, subtropical and maritime environments, making them very good complements for tree rings in terms of spatial coverage. Corals also provide a uniform window of climate information because they can be precisely dated (annually and seasonally) and their environment can be sampled continually over the period of a year. Its ability to determine past climate events in the ocean makes it useful in prediction analysis of the ocean climate.

VARVED LAKES AND OCEAN SEDIMENTS RECORDS

The accumulation of sediments at the floors of Oceans and Lakes serves as a good source of past climate information. Similar to ice cores, information on past climates are imbedded within the layers of the ocean sediments. Scientist drill cores into the billions of tons of inorganic sediments that have accumulated overtime from the ocean and lake floors and examine the properties to determine past climate. Varved Lakes (VL) are important complements to proxies like tree rings as they provide climate information in high latitude regions where such proxies are limited or unavailable (Jones and Mann, 2004, IPCC, 2007). The deposition of sediments in Varved Lakes are controlled by seasonal precipitation and temperature changes, the amount of melted water discharge and sediment load into a bounded glacial lake are influenced by precipitation and temperature pattern. Varved Lakes are important in determining the climate temperature at the time the layer was formed.

OTHER PROXIES

There are certain other proxies that are still under development or seldomly used because of low resolution or the technicalities involved in their calibration process. Some of these other proxy sources include, but are not limited to; isotopes form molluscs, glacial evidence and boreholes.

TIMING OF PROXIES

As these climate proxies are used to define past climate with particular attention to year and season of occurrence, it is important to know how the time frame for each of these proxies are determined. For proxies like tree rings and corals that develop their layers annually, the rings are simply counted to determine the exact year of occurrence. In other cases, radiometric dating is used to determine time. Radioactive elements within the proxies reveals certain properties that can be used to estimate the time of occurrence, however in certain cases for very old proxies, some of the older radioactive elements might have decayed producing different proportion of elements which may not be found in newer proxies (Osama *et al*, 2020).

b. Strength and weakness of datasets used in climate extremes studies

In the study of climate extreme systems (*I believe would have been explained prior this section*), observational and gridded datasets have provided quality representation in spatial and temporal understanding of our environment. Atmospheric variables can be monitored consistently over time providing past and present understanding of climate systems. Today's understanding about the climate is developed from continuous observations and records of the evolution of past systems. These records of observations are referred to as historical climate datasets and are vital towards the accurate understanding of the dynamics of climate systems. Climate datasets are continuous with past records providing understanding for present day knowledge, and current observations as basis for the future.

More data means more information and understanding can be harnessed. This is important to climate scientists and forecaster for accurate prediction of climate systems on a short- or long-term. With great variability of observed climate systems, the quality of our understanding depends on how much data are available to be accessed.

Slight change in the severity or frequency of occurrence in an extreme climate event could have profound effect on the environment. The changes in wet spells for instance would have tremendous effect on agriculture. The extent of these changes is quantified by the amount and quality of dataset used. Hence, it is important we examine the strengths and shortcomings of the type of datasets involved.

2.2.1 Gridded datasets: Observations, Reanalysis, Satellite and Climate models

These datasets are most commonly used in present day climate studies. Unlike station data, gridded datasets are measurements or data interpolated at a regular or homogenous grid. These datasets are either reanalyses products, derived through remote sensing or interpolated station observations. Usually they extend over the whole globe or a limited area (regional). Gridded data involves the grid representation or interpolation of multiple observation networks. Thus, they provide an extensive view of the regional extent over the area of interest, and giving

information of stations with missing or no data. Measurements from satellites are usually processed in this format extending over the region of satellite coverage.

Reanalyses or Climate models' output are represented in this data type. Although, dependent on their resolution scales, they provide adequately high spatial representation of climate data. High or fine resolution climate data are provided at smaller grid sizes or intervals, and are more informative in regional studies. For example, precipitation retrievals from the Global Precipitation Measurements (GPM) is provided on a 10km x 10km grid interval. Coarse or lower resolution datasets are less informative for regional purposes and more suitable for global studies. An extended range of climate data are available within this grid, ranging from 0.5° to about 1° resolution. For example, the National Oceanic and Atmospheric Administration (NOAA), the European Centre for Medium-Range Weather Forecasts (ECMWF), or the National Center for Atmospheric Research (NCAR) provides continental or global data extending over a large domain and a period of time. However, they are of limited usage in regional or small-area studies e.g. precipitation over a small catchment area or surface temperature in heat island analysis.

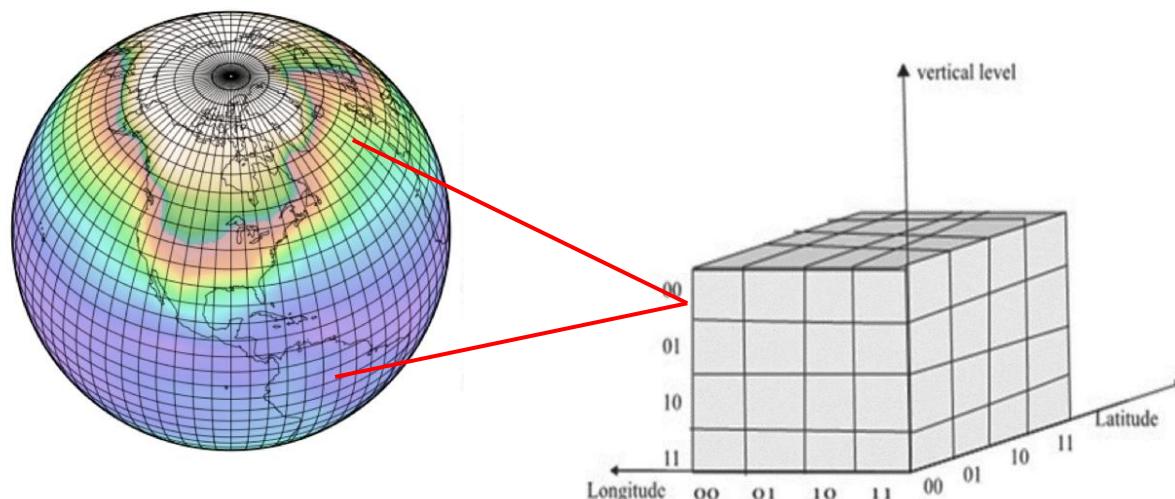


Figure 1: A three-dimensional global grid system (source: <https://www.nccs.nasa.gov/services/climate-data-services>; Varalakshmi et al. 2020)

Gridded climate data are quality controlled and suitable for use either in data assimilation or in statistical/physical process studies. These datasets have undergone multiple screening processes through sufficient analysis for biases, outliers, or error characterization before accessed by the users.

Shortcomings of Gridded dataset in Climate Extreme Studies

Gridded data from satellites and reanalyses are not without their limitations. One of which is their lack of long historical records. For example, precipitation retrievals from satellite measurements only extends back to the 1970s, and there are significant biases existing in the data (Gerstner and Heinemann 2008). As the evaluation of extremes require long historical

records to identify significant and consistent factors attributed to these extremes. In analyzing a one-in-ten years extreme e.g. a drought event, twenty to thirty years data may be insufficient to reach a logical scientific conclusion.

This type of dataset is not also without biases (underestimation or overestimation) in representation of climate extremes. For gridded observation networks, the quality of gridded product is related to the number of contributing meteorological stations (Haylock et al. 2008). More observation networks suggest better interpolation of observed variable across the grid region. Also, complex topography or terrains are associated with unique weather systems varying from surrounding plains or regions, thereby poor representation of mountains or cliffs in the observation stations would result in a gridded product with large biases. Additionally, when examining extreme rainfall data, there are dangers of over-smoothing precipitation data when there are few stations used in the analysis. (Hofstra et al. 2010)

Climate models or reanalyses products are associated also with significant biases. Reanalyses data from the National Centers for Environmental Protection/National Center for Atmospheric Research (NCEP/NCAR) have significant underestimation in precipitation extremes (Hanson et al. 2007). Reanalyses data generally differ due to disparities in assimilated observational data and methods, the configurations of boundary layer processes, and the physics schemes. These contributes to the level of uncertainties in the output product, which may be considerably large in climatic extremes analysis. For instance, discrepancies between reanalyses for some climate extreme indices, such as frost days in some regions, are sometimes as large as the typical inter-model spread of the Coupled Model Intercomparison Project ensembles, (Sillmann et al., 2013). The European Centre for Medium-Range Weather Forecasts (ECMWF) Interim reanalysis, the fifth generation of ECMWF atmospheric reanalysis (ERA-5) and the Department of Energy (DOE) Reanalysis 2 are most common reanalyses data in present-day climate studies. A major limitation in earlier versions of ECMWF products are large biases in precipitation estimates., while the newer version (ERA-5) has major improvements in circulation patterns hence reducing these rainfall biases (Nogueira, 2017; 2020).

The use of the ensemble average from multiple climate models have been adopted lately in most scientific studies. The configuration and parameterization of schemes differs in models. The performance of these models also varies across regions affected by topography and the dynamics of the climate. The result is in overestimation or underestimation of extremes, varying in degrees with different GCMs or RCMs. The ensemble Mean of these models is a way to reduce these biases. To estimate the mean ensemble in climate models requires all models to be represented on equal grid resolutions.

Systematic biases are other limitations created through the use of gridded climate models in analyses of climate extremes. For example, since models or reanalyses uses grid averages, differences in hot and cold extremes are usually smaller across nearest grid locations as compared to real time observations.

Additionally, non-stationarity is another limitation associated with some reanalysis products. As reanalyses data combine observational datasets from differing sources on a long-term, increasing variations in mean or variances affect the overall trend. Temporal inhomogeneities

due to changing assimilated observations suggest reanalysis products are unsuitable for long term climate analysis. The sensitivity of climate extreme analysis suggests these could contribute to the uncertainties associated to its studies.

The highlighted issues above are the major limitations or weaknesses associated with the use of gridded climate data in extreme climate scenario.

Advantages of Gridded dataset in climate extreme studies

The wide acceptability of gridded datasets in climate studies is met with its ability to provide a wide or regional spatial information at a glance. Large scale regional studies are quite impossible with station observation datasets. Hence, variability of systems responsible for the evolution of climate extremes can be easily monitored and traced to source regions with gridded datasets.

Another strength of gridded data in extreme climate studies is their ability in providing future climate projections. The effects of the changing climate suggest that while it is important to understand past or historical events, the knowledge of the future is more relevant to the state-of-art science. Station observations are available at a fine resolution but only provide historical observations at best. The continuous improvement in climate models and data assimilation methods means they can become reliable in the projections of climate extremes. New groups of models have incorporated finer spatial resolution, new physical processes and biogeochemical cycles. For example, most models in the recent version of the Coupled Model Intercomparison Project-Phase 6 (CMIP6) have an improved climate sensitivity, contributing to higher warming projections up to 0.4°C as compared to similar scenario in the previous version.

Gridded data also offers some advantages through availability of data in non-observed regions. The use of station or point observations means data is only made available at specific instrumented locations.

Regional or zonal analysis is simplified with gridded datasets. Unlike point measurements, averaging data over region is straight forward in gridded dataset.

2.2.2 Station or Point Observations

Station observations provide more accurate historical record of climate events, thereby improved results in climate extremes studies. They are helpful in providing significant details in daily extreme analysis. Unlike climate extreme, daily extremes have a shorter temporal span lasting between hours to days. Example is an abnormal increased (decreased) temperature exceeding normal daily averages, generally referred to as heat (cold) waves.

Shortcomings of Station Observations

One of the major limitations of station observation is the coverage of less spatial extent. They provide interesting results for microclimate studies, and cannot be applied in monitoring synoptic or mesoscale events.

In hydrological studies, particularly in stream flow experiments or simulations, instrumental measurements of meteorological variables are of great concern. Most especially, precipitation measurements are identified with different error sources including aerodynamic, wetting, evaporation, splash in and out, and blowing and drifting snow factors, thus leading to uncertainty in precipitation estimates (Taskinen and Söderholm 2016). Microclimatic variations (for instance a local storm) can also be challenging in the representativeness of measured precipitation estimates (Orlowsky and Seneviratne, 2014).

Temperature measurement are more sensitive to the environment. Error arises from thermometer exposure to solar radiation and the environment. Station observations can also be subjected to bad observer practices or poor data processing. Ultimately, poor-quality data resulting from these errors can lead to misinformation or incorrect model calibration (Beven and Westerberg 2011).

Advantages/Strengths of Station Observations

The sensitivity of temperature measurements means stations observations are most advisable to be used in local scale studies. Heat waves monitoring can vary within the local climate, and are subject to environment conditioning. Hence, point measurements can provide instantaneous monitoring of extreme weather events.

Station observations have less uncertainties resulting computational errors, and provide improved accuracy in historical monitoring of climate extremes.

They are also available in fine resolutions are very useful in microclimate studies.

c. The Climate System: A data science perspective

Climate has a significant impact on life on earth because it plays an important role in the daily experience of human and it is essential for health, food production and well-being. However, a report presented scientific evidence that human activities are already influencing the climate (IPCC, 2013). If we wish to understand, detect and predict the impact of human on climate, we need to understand the system that determines the climate of the earth and of the processes that lead to climate change (Baede et al., 2018). Global climate change has emerged as the greatest environmental challenge of our era (21st century). Hence, understanding our changing world has impelled many researchers from different fields of science to tackle complicated research questions. The climate change research community now faces the daunting task of disseminating massive amounts of information about possible future climates under varying scenarios to a large audience (Pickard et al., 2015). They also need to make the data readily accessible so that it can be used by scientists in other research fields.

In understanding the climate of our planetary system (Earth) and its variations, and possibly predicting the changes of the climate influenced by human activities, emphasis should be on those factors and components that determine the climate. It is important we must understand the *climate system*, the complicated and interactive system consisting of five major components, including the atmosphere, the hydrosphere, the cryosphere, the lithosphere and the biosphere (Faghmous et al., 2014; Baede et al., 2018). Directly or indirectly, all the components are affected as a result of the global climate change. Hence, an adequate response requires relating changes in the global climate to their impacts on local communities, such as flood, severe droughts, wildfires, and hurricanes. However, this attempt faces significant challenges because existing climate-model for studying these changes don't resolve many climate change impact phenomena at spatiotemporal scales relevant to policymakers, community leaders, and other stakeholders (Faghmous et al., 2014; Faghmous and Kumar, 2014; Hassani and Huang, 2019).

Based on these limitations and the urgency of climate change, this necessitates the opportunity for the data-driven methods to fill up knowledge gaps related to climate change and its societal impacts. Climate science is a field focused on studying large-scale changes across the components of climate system over long and short temporal periods and is becoming an increasingly ripe domain for significant data-science contributions as data from earth orbiting satellites, climate-model simulations, and paleoclimate records have been growing exponentially and will continue to do so in the next decade (Faghmous et al., 2014). However, Faghmous and Kumar (2014a) identify three major factors that could possibly slowed the progress. Foremost, the data that climate science employs violate many of the assumptions and practices held in traditional data science. For instance, the majority of climate data are organized in a spatiotemporal grid. Intrinsically, the data are auto-correlated where regions in spatial or temporal proximity tend to be highly related. Therefore, any methods that impose independence assumptions among data points will have limited practicality with such data. Secondly, they observed that the field of data science has historically focused more on certain tasks and evaluation metrics (Langley, 2011; Faghmous and Kumar, 2014a; Jagadish, 2015) that are not applicable to some of climate science's biggest needs (Hassani and Huang, 2019). Lastly, though is only a matter of time, climate science, its data, and challenges have not been adequately exposed to the broader data science community until recently.

Table 1: List of Climate Data Sources with Temporal Coverage of Each Data Source

	Key Strength	Key weakness
Climate model	Ability to run forward simulations	Relies solely on physics
<i>In situ</i> observation	Direct observations	Spatial Bias
Satellite	Global coverage	Lack of continuity as missions last on average 5 years
Paleoclimate	Ability to use proxy data to infer pre-industrial climate trends	Techniques to analyze such data are still evolving

Source: Faghmous and Kumar (2014a)

d. Data challenges and continuously changing data

The use of climate data for applications and research globally has been scanty due to the unavailability of and access to climate data is very limited. In many regions of the world especially, developing countries, weather stations are sparse and their number has been declining (Dinku, 2019). Besides, the distribution of existing stations is uneven, with most located along major roads and other disturbances. Sparsity of climate data can refer to the absence of data required to generate useful climate information to perform meaningful analysis and inform climate-resilient growth. Data sparsity usually refers to the situation in which climate data is not usable or accessible. Although this problem is prevalent globally, it is especially prevalent in the developing nations, particularly in areas with difficult and remote geographies, where conflict and data investment are a relatively low priority (Hunziker et al., 2017; Hunziker et al., 2018). The meteorological or climatological agencies in any area are the primary sources of climate observations. However, observation networks are insufficient in different regions of the world, especially in developing countries, with the number and quality of weather stations declining in many parts of the continents (Dinku et al., 2017; Parker et al., 2011). Moreover, as most of the current weather stations have been found to be unevenly distributed, with most of the stations clustered in cities and towns. As a result, climate data may not be available in rural areas where it can be verified that this data is most needed, with very few stations in the forested and desert regions (Rotenberg and Yakir, 2010). In different applications, such as location-based and sensor services, climate data is continually changing (Xia et al., 2005). The issue of using data for the efficient assessment of environmental components is increasingly relevant. Conventional indexes may have suffered from continuously changing climate data due to the constantly changing nature of climate data, which may contribute to poor performance (Xia et al., 2005).

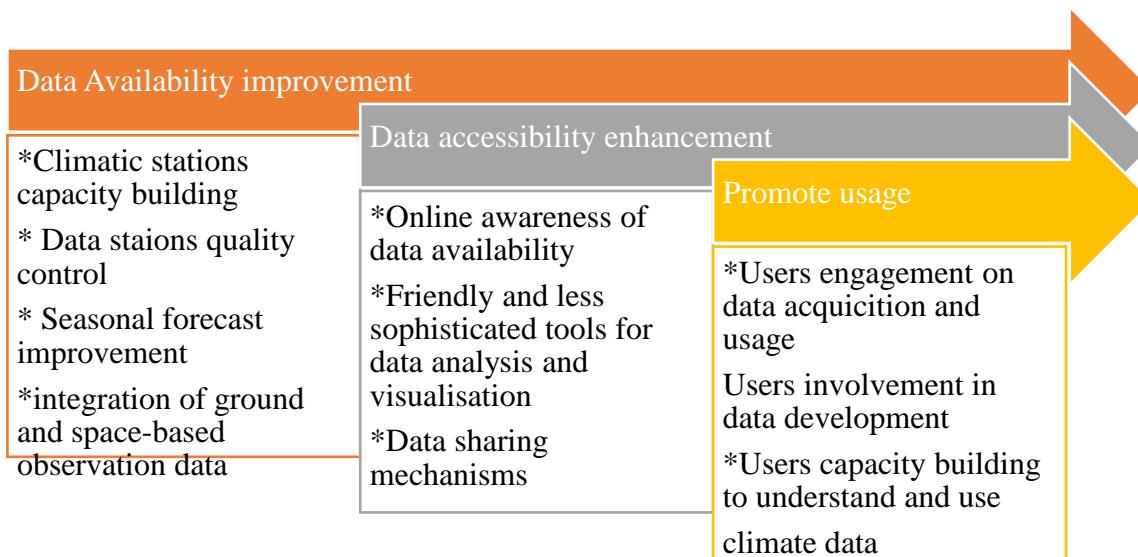
Sparse station distribution is not the only challenge; the number of observation stations has also been declining for decades in many regions. This decline may be attributed to some factors:

the first one could be that the data are available but may not have been provided to the appropriate institutions, and the second factor is an actual decline in observation, the third one is inadequate tools and facilities (lack of finance). For example, between 1971 and 2001, the average number of active stations in Madagascar declined from over 400 to under 50 (Dinku, 2019). This is a very serious loss and challenge to the use of climate data for research and different applications in the region. The decline in investing in climate infrastructure is also major obstacle to the operation and maintenance of climate observation networks and related infrastructure, especially in low-income countries, for many climate services. This may be due in part to difficulties in articulating the value added by meteorological agencies (Rogers and Tsirkunov, 2010; Bouwer et al., 2014) and often to a lack of awareness of the benefits of climate observations for growth (Hansen et al., 2007; Bryan et al., 2009; Roncoli et al., 2009).

To improve the availability, access, and use of climate data and derived information products for research and applications, efforts to be made which will represent Data usage improvement components (Table 1), which are data availability improvement, data accessibility enhancement, promote usage of climate data and information products. Rising types and volumes of climate data alone are a major challenge for the climate science community and its funding institutions. In order to generate, format, record and distribute all these data, institutional capacity must exist, while at the same time a much larger group of diverse users applauds for access, understanding and use of climate data. These include a rising number of scientists (across different fields) and decision-makers with real resources, livelihoods, and even lives at stake in society (resource managers, farmers, public health officials, and others). Key users, as well as their constituents in the general public, also include those with public responsibilities who may support and understand the decisions that are taken on their behalf (Overpeck et al., 2011; Faghmous and Kumar, 2014). As a result, climate scientists not only have to share data with each other, but also have to meet a growing duty to promote access to

data for others outside their community, and to respond to the wider users in order to ensure that data is as useful as possible (Table 1).

Table 1. Data usage improvement components



In order to mitigate the challenges of data availability and access, different efforts are required. These include the interpolation of existing observations from stations and the use of proxies, such as forecasts of satellite rainfall and products for climate model reanalysis. A new approach should aim at improving the availability, accuracy and access of data by combining quality-controlled station data from the entire national observation network with proxies such as satellite climate data estimates and climate model reanalysis products is important. This will include comprehensive climate data and targeted information products that are directly applicable to the needs of decision-makers at different levels, allowing a number of users to leverage past, current and future climate information. Climate scientists and other types of scientists who work effectively at the interface between science and applications can increasingly interact closely with climate stakeholders in society, and also with collaborative knowledge generation (Held, 2004; Knight and Jäger, 2009). Together with the interdisciplinary research community that supports them, these stakeholders are the users that would drive the future climate data project.

e. Short observational record and evaluation challenges

Climate observations are sourced from numerous meteorological and related observational networks and systems throughout the world. Narama et al 2010, stipulated in

their study that to describe the whole climate system, we need to collect observations of the atmosphere, ocean and land-based systems as it is these particular entities that govern and supply the data for making deductions and sensings with regards to climate fluctuations and consistencies. This is commonsensical and can be attuned to the importance and involvement of water bodies in the overrunning of affairs that decide how Climate can be observed and how the challenges can be surmounted. Also, Lenderink and Meijgaard (2018) tallies increases in hourly precipitation extremes beyond expectations from temperature changes as core observational parameters for gauging climate change. However, not all climate observations are the same and some are considered of higher quality than others. Mitrovica 2015, made it known as a corroboratory inference from his research that therefore, a climate observation should be associated – either directly or indirectly – with a set of data (Metadata) describing the conditions in which the climate observations were collected and that will provide users with information, often implicitly, on how the observations should be interpreted and used. This is straightforward and is tech-savvy, comprehensible and accessible. According to the Journal of Geophysical research, 2019, assessing the impacts of our changing climate can be and is scientifically challenging.

Historically, only a limited amount of information about past climate change and its future impacts has been available at national level, while approaches to the science itself have varied between countries. The statistical survey of Science Direct, 2019 reveals that there are over 11,000 weather stations around the world measuring land, air and sea temperatures, as well as satellites, ships and aircraft that also take measurements. There is the place of gears, facilities and funding, and then there is the place of qualified personnel and intellectual verbosity.

Evaluating challenges that come with observational studies can be daunting. (Lambeck et al, 2010) argued that the influence of geophysical modelling on the accuracy of observational studies can be so compelling that concise results don't get released until longer timescales are spent. While the world has agreed on the need to tackle climate change for sustainable development, it is critical to provide evidence-based analysis of past experiences and ongoing innovations to shed light on how we might enhance the effectiveness and efficiency of actions at various levels. Kjellstrom et al, 2009 posited that thorough and credible evaluations help us identify what works, for whom, when and where and under what circumstances in order to mitigate climate change, achieve astute situations for the society, the economy and the environment, reduce risk and increase resilience in the face of changing climate conditions. The onus lies on the climatologist to be innovative especially in a resource-limited setting.

REFERENCES

Baede, A. P. M., Ahlonsou, E., Ding, Y., & Schimel, D. (2018). The Climate System : an Overview.

Bouwer, L. M., Papyrakis, E., Poussin, J., Pfurtscheller, C., & Thielen, A. H. (2014). The costing of measures for natural hazard mitigation in Europe. *Natural Hazards Review*, 15(4), 04014010.

Bryan, E., Deressa, T. T., Gbetibouo, G. A., & Ringler, C. (2009). Adaptation to climate change in Ethiopia and South Africa: options and constraints. *Environmental science & policy*, 12(4), 413-426.

Dinku, T. (2019). Challenges with availability and quality of climate data in Africa. In *Extreme hydrology and climate variability* (pp. 71-80). Elsevier.

Dinku, T., Thomson, M. C., Cousin, R., del Corral, J., Ceccato, P., Hansen, J., & Connor, S. J. (2018). Enhancing national climate services (ENACTS) for development in Africa. *Climate and Development*, 10(7), 664-672.

Faghmous, J. H., Banerjee, A., Shekhar, S., Steinbach, M., Samatova, N., & State, N. C. (2014). *Theory-Guided Data Science for Climate Change*. 74–78.

Faghmous, J. H., & Kumar, V. (2014a). A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. *Big Data*, 2(3), 155–163.
<https://doi.org/10.1089/big.2014.0026>

Faghmous, J. H., & Kumar, V. (2014b). Spatio-temporal Data Mining for Climate Data: Advances , Challenges, and Opportunities. In *Data Mining and Knowledge Discovery* for Big Data. <https://doi.org/10.1007/978-3-642-40837-3>

Gerstner, E. and Heinemann, G. (2008). Real-time areal precipitation determination from radar by means of statistical objective analysis. *Journal of Hydrology*, 352: 296–308, DOI:10.1016/j.jhydrol.2008.01.016.

Hansen, J. W., Baethgen, W. E., Osgood, D. E., Ceccato, P. N., & Ngugi, R. K. (2007). Innovations in climate risk management: protecting and building rural livelihoods in a variable and changing climate.

Hassani, H., & Huang, X. (2019). *Big Data and Climate Change*. 1–17.
<https://doi.org/10.3390/bdcc3010012>

Hanson, C.E., Palutikof, J.P., Livermore, M.T.J. et al. (2007) Modelling the impact of climate extremes: an overview of the MICE project. *Climatic Change*, 81, 163–177. DOI:10.1007/s10584-006-9230-3

Haylock, M., Hofstra, N., Klein Tank, A., Klok, E., Jones, P. and New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *Journal of Geophysical Research*, 113: D20119, DOI: 10.1029/2008JD010201.

Held, I. M. (2005). The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, 86(11), 1609-1614.

Hofstra, N., New, M. and McSweeney, C. (2010). The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Climate Dynamics*, 35: 841–858. DOI:10.1007/s00382-009-0698-1

Hunziker, S., Brönnimann, S., Calle, J., Moreno, I., Andrade, M., Ticona, L., ... & Lavado-Casimiro, W. (2018). Effects of undetected data quality issues on climatological analyses. *Climate of the Past*, 14(1), 1-20.

Hunziker, S., Gubler, S., Calle, J., Moreno, I., Andrade, M., Velarde, F., ... & Brönnimann, S. (2017). Identifying, attributing, and overcoming common data quality issues of manned station observations. *International journal of climatology*, 37(11), 4131-4145.

IPCC. (2013). Climate change 2013: The physical science basis. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Contribution of Working Group I to the Fifth Assessment Report of the. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (p. 1535). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Intergovernmental Panel on Climate Change 2007. Climate Change 2007 - The Physical Science Basis Contribution of Working Group I to the Fourth Assessment Report of the IPCC (ISBN 978 0521 88009-1 Hardback; 978 0521 70596-7 Paperback)

Jagadish, H. V. (2015). Big Data and Science : Myths and Reality. *Big Data Research*, 2(2), 49–52. <https://doi.org/10.1016/j.bdr.2015.01.005>

Jones, P. D., and M. E. Mann (2004), Climate over past millennia, *Rev. Geophys.*, 42,

RG2002, doi:10.1029/2003RG000143.

Knight, C. G., & Jäger, J. (2009). *Integrated Regional Assessment of global climate change*. Cambridge University Press.

Langley, P. (2011). The changing science of machine learning. *Mach Learn*, 82, 275–279.

Pickard, B. R., Baynes, J., Mehaffey, M., & Neale, A. C. (2015). *Translating Big Data into Big Climate Ideas*. April.

NOAA. What Are “Proxy” Data? <https://www.ncdc.noaa.gov/news/what-are-proxy-data>

Nogueira M. (2017). Exploring the links in monthly to decadal variability of the atmospheric water balance over the wettest regions in ERA-20C. *Journal of Geophysical Research - Atmosphere*, 122. DOI: 10.1002/2017JD027012

Nogueira, M. (2020). Inter-comparison of ERA-5, ERA-Interim and GPCP rainfall over the last 40 years: process-based analysis of systematic and random differences. *Journal of Hydrology*, DOI:10.1016/j.jhydrol.2020.124632

Osama R. Shaltami, Jorge Artola, Artemisia Cordoba, Olivia Arconada, George D. Siasia, Maribel R. Fernandez, Larsen Jorgensen, Christian Elkjaer, Fares F. Fares, Farag M. EL Oshebi, Hwedi Errishi and Elena Maceda, 2020; Paleoclimatology – A Review. 3rd International Symposium on Paleoclimatology, Federal University of Goiás, Brazil.

Orlowsky B. and Seneviratne, S. (2014). On the spatial representativeness of temporal dynamics at European weather stations. *International Journal of Climatology*, 34, 3154–3160. DOI:10.1002/joc.3903

Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. *science*, 331(6018), 700-702.

Parker, D., Good, E., & Chadwick, R. (2011). *Reviews of observational data available over Africa for monitoring, attribution and forecast evaluation* (p. 62). Met Office.

Rotenberg, E., & Yakir, D. (2010). Contribution of semi-arid forests to the climate system. *Science*, 327(5964), 451-454.

Rogers, D., & Tsirkunov, V. (2011). Costs and benefits of early warning systems. *Global assessment rep.*

Roncoli, C., Jost, C., Kirshen, P., Sanon, M., Ingram, K. T., Woodin, M., ... & Hoogenboom, G. (2009). From accessing to assessing forecasts: an end-to-end study of participatory climate forecast dissemination in Burkina Faso (West Africa). *Climatic Change*, 92(3), 433-460.

Sillmann, J., Kharin, V., Zhang, X., Zwiers, F. and Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research -Atmosphere*, 118, 1716–1733. DOI:10.1002/jgrd.50203

Taskinen, A. and Söderholm, K. (2016). Operational correction of daily precipitation Measurements in Finland. *Boreal Environment Research*, 21, 1–24. <https://helda.helsinki.fi/bitstream/handle/10138/225261/ber21-1-2-001.pdf?sequence=1>

Varalakshmi, M., Kesarkar, A. P., & Lopez, D. (2020). High-performance implementation of a two-bit geohash coding technique for nearest neighbor search. *Concurrency and Computation-practice and Experience*, 33:e6029, DOI:10.1002/cpe.6029

Xia, Y., Prabhakar, S., Lei, S., Cheng, R., & Shah, R. (2005, March). Indexing continuously changing data with mean-variance tree. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1125-1132).