Water Resources Management (2022) 36:4019–4038
https://doi.org/10.1007/s11269-022-03231-z

Comparative Evaluation of Water Resource Data Policy Inventories Towards the Improvement of East African Climate and Water Data Infrastructure

Victoria M. Garibay[1] · Margaret W. Gitau[1] · Victor Kongo[2] · James Kisekka[3] · Daniel Moriasi[4]

Received: 22 October 2021 / Accepted: 23 June 2022 / Published online: 29 July 2022

Abstract

The recognized challenge of freely accessing climate and water data in East Africa poses a problem in undertaking relevant analytical studies and making informed water resources management decisions in the region. This study seeks to understand the defining characteristics of policies and distribution infrastructure, in the context of meteorological, water quantity, and water quality data, that determine whether or not a user will be able to freely and readily access existing data. An analysis was developed to quantify the information contained in legislation, official documents and websites, and similar textual resources from the study region and elsewhere to establish commonalities, potential trends, and patterns in the documentation behind data streams culminating successfully in a portal or database accessible by the public. A quantitative analysis was applied to discern overall patterns in what constitutes effective policy and to diagnose where there may be impediments in the path between data collection and its application. Generally, the foundational elements present in the documentation pertaining to most accessible data streams represented are: (1) known organization in charge of that data type; (2) known location where this data would be stored; (3) defined data collection format; and (4) commitment to a plan for making data available to potential users. Examination of overlap between elements absent in unsuccessful data streams and present in successful data streams suggests that those without a documented commitment to making data available online rarely result in a functioning, accessible portal and vice versa. Amongst other findings, this knowledge has the potential to contribute towards the development and refinement of policies so that more emphasis is placed on openness and access, leading to informed decision-making and management of water resources.

✉ Margaret W. Gitau
mgitau@purdue.edu

[1] Purdue University, West Lafayette, IN, USA

[2] Global Water Partnership Tanzania, Dar es Salaam, Tanzania

[3] RAIN (Aidenvironment), Kampala, Uganda

[4] USDA-ARS Grazinglands Research Laboratory, El Reno, OK, USA

1 Introduction

Meteorology, water quantity, and water quality data play critical roles in water resources planning for rural and urban applications. Without essential data infrastructure, both physical (e.g., monitoring stations) and intangible (e.g., databases, processes, policies, and guidelines), a national service cannot generate and effectively disseminate data and related analytics (Mason et al. 2015). Additional hurdles arise as technology evolves, thereby increasing the volume and complexity of data, and overwhelming existing data infrastructure (Overpeck et al. 2011). In regions of data scarcity, there is uncertainty regarding from where challenges to data collection, accumulation, and distribution processes stem and what their direct causes are. This uncertainty makes it difficult to determine what changes or improvements are needed to facilitate and streamline the process as a whole. Thus, the identification of underlying causes becomes a natural next step towards increased understanding and diagnosis of the issue.

Data mining is useful in identifying and evaluating conceptual linkages (Tian et al. 2020). The collection of data from textual sources takes many forms, including key phrase searches, fuzzy functions, phrase extraction, and others (Aman et al. 2021; Bui et al. 2016; Gitau et al. 2005; Turksen and Celikyilmaz 2010; Wang et al. 2017; Zaeem et al. 2018). Many of these methods are performed on large batches of documents, mining key information or data which is then condensed into a database. Examples include aggregating phrases to form summaries of articles or books (Bui et al. 2016; Wang et al. 2017) and isolating the quantitative results of a predefined set of practices from multiple studies (Gitau et al. 2005).

The policies and official resources resulting in successful data streams can serve as a benchmark to determine what information is missing from, or might necessitate revisions of, existing policies in a study region. Specific elements, selected based on preliminary examination of the documents with consideration for the principals and examples of open data as defined by The World Bank (2019), can be mined from the texts to form a more complete picture of their strengths and shortcomings.

In the East African countries of Kenya, Tanzania, and Uganda, the recognized lack of reliable, freely accessible data poses a problem for development and application of tools such as hydrologic and water quality models and other analyses that can be used to inform water resources management decisions (Garibay et al. 2021; Gebrechorkos et al. 2018; Näschen et al. 2018; Mogaka et al. 2006). This issue extends to many data types including precipitation, solar radiation, wind speed, temperature, streamflow, and pollutant concentration data which play roles in scientific evaluation and projection to address water budgets, pollution regulation, flood risk, and many additional concerns.

Global datasets obtained via remote sensing and reanalysis methods can be a valuable resource in East Africa and other data scarce regions (Garibay et al. 2021; Gebrechorkos et al. 2018; Lakew et al. 2020). These highly advanced, approximations are being used as an alternative to measured data; however, it is important to consider the uncertainty inherent in these methods (Lakew et al. 2020); instances of bias have been noted in the representation of precipitation and temperature for East Africa, making measured data a necessary input in assessing these differences so that they can be corrected or noted in evaluations (Garibay et al. 2021; Gebrechorkos et al. 2018; Lakew et al. 2020).

In this study, an analysis of the commonalities and differences in country-specific policies and documentation from within and outside the study region was conducted to determine what aspects led to successful data streams—defined as when a non-affiliated user can freely access and download available data records. Specifically, the study: (1) investigated potential sources of disruption between collection and effective use of water and climate data; (2) compared and contrasted data policies and policy documentation from different countries and regions; and, (3) made recommendations on data infrastructure and policy enhancements that could clarify documentation and improve data accessibility.

Similar to a method used to analyze privacy policies (Zaeem et al. 2018), a survey was taken of relevant documents to provide a deeper understanding of the functionality of policies in the primary study area and guidelines relating to the stewardship of data pertinent to informing water resources management. Unlike the multi-level responses evaluated in the privacy policy analysis (Zaeem et al. 2018), the survey was styled to retrieve binary responses. A quantitative analysis method was developed to aggregate these responses based on the goal of scientifically assessing the qualitative information contained within the documents and resources examined. First presented is the methodology used in the analysis. Next, the results of this analysis are reported. Finally, these results are used to inform discussions by highlighting important aspects, bottlenecks, and avenues for further consideration so that recommendations can be made for future policy improvements in the region.

2 Methodology

2.1 Data Sources

Examples of official resources surveyed include legislative acts, policies, guidelines, procedures, strategies, mission statements, and webpages from governing entities that were publicly available online (Table 1). Items produced by non-governmental organizations and sources such as news, books, and journal articles were not considered in the scope of this analysis. To accommodate the comparisons necessary for this study, the documents and procedures examined went beyond the primary study region of Kenya, Tanzania, and Uganda (Fig. 1). The other countries chosen for comparison were Burundi, Ethiopia, and Rwanda—for a more inclusive overview of East Africa—and Canada, Germany, South Africa, and the United States, selected for their reputations for open data availability. Over 50 documents were surveyed, distributed as 3–7 resources from each of the 10 countries (Table 1, Online Resource 1). The number of documents considered for each country varied based on documentation style and availability of information to the public through online portals. This list may not reflect all resources currently available due to rapid changes in internet resources, with some documents thought to be available not found and possible introduction of new documents beyond the time documents were accessed for this study. Likewise, a listed resource may have since been removed from public access or amended.

2.2 Data Mining Strategy Development

Due to their verbal nature, legislation, mission statements, and similar documents teem with qualitative data which is highly relevant to assessing the strength and robustness of an approach to the national management of water resources data. However, in its qualitative form, it can be challenging to analyze from an objective, scientific standpoint. Therefore, a

Table 1 Resources listed by country including type and date surveyed and corresponding citation ID in Online Resource 1

Country	Name of Resource	Resource Type	Date Surveyed	ID
Burundi	Stratégie Nationale de l'Eau	Official Document	29-Apr-21	29
	Code de l'Eau au Burundi	Legal Document	29-Apr-21	13
	IGEBU Institut Géographique du Burundi (http://igebu.gov.bi/)*	Official Website	29-Apr-21	26
Canada	Historical Climate Data (https://climate.weather.gc.ca/)	Official Website	23-Apr-21	11
	Water Level and Flow (https://wateroffice.ec.gc.ca/)	Official Website	23-Apr-21	12
	Freshwater quality monitoring: online data (https://www.canada.ca/en/environment-climate-change/services/freshwater-quality-monitoring/online-data.html)	Official Website	23-Apr-21	10
	Water Survey of Canada Official Products and Outputs	Official Document	23-Apr-21	31
	Canada Water Act	Legal Document	23-Apr-21	9
	Canadian Environmental Sustainability Indicators: Water Quality in Canadian Rivers	Official Document	25-Apr-21	8
Ethiopia	Ministry of Water, Irrigation and Electricity (http://mowie.gov.et/)	Official Website	29-Apr-21	37
	Ethiopian Water Sector Strategy	Official Document	30-Apr-21	33
	NMA: National Meteorology Agency (http://www.ethiomet.gov.et/)*	Official Website	30-Apr-21	43
Germany	Die deutschen Klimabeobachtungssysteme	Official Document	4-Dec-20	5
	Deutscher Wetterdienst Act (DWD Act)	Legal Document	11-Dec-20	4
	Wasserhaushaltsgesetz – WHG (Water Resources Act)	Legal Document	11-Dec-20	15
	Abwasserabgabengesetz - AbwAG (Wastewater Charges Act)	Legal Document	12-Dec-20	14
	Deutscher Wetterdienst (https://www.dwd.de/)	Official Website	3-Jan-21	6
	bfg: Bundesanstalt fur Gewässerkunde (https://www.bafg.de/)	Official Website	3-Jan-21	1

quantitative analysis method was developed and applied to observe overall patterns in what constitutes effective documentation and to diagnose where there may be impediments in the path between data collection and effective data use.

Document data-mining was accomplished using a survey, with a broad focus on meteorology, water quantity, and water quality as the three overarching data types of interest. These three are hereafter referred to as themes. Each theme was comprised of nine defined elements—namely Format, Instrumentation, Digitization, Destination, Service Provision, Responsibility, Chain of Command, Functionality, and Accessibility—across four categories, Data Collection, Data Curation, Governance, and Performance (Table 2). The selection of elements was based on a preliminary examination of documents from multiple coun-

Table 1 (continued)

Country	Name of Resource	Resource Type	Date Surveyed	ID
Kenya	The Water Act 2016	Legal Document	8-Dec-20	16
	KMD Meteorology Policy	Legal Document	8-Dec-20	27
	Kenya Meteorology Bill, 2019	Legal Document	8-Dec-20	17
	Water Resources Authority Strategic Plan 2018–2022	Official Document	8-Dec-20	52
	Environmental Management and Co-Ordination (Water Quality) Regulations, 2006	Legal Document	8-Dec-20	30
	Water Resources Authority (https://wra.go.ke/)	Official Website	19-Dec-20	53
	Kenya Meteorological Department (https://www.meteo.go.ke/)	Official Website	19-Dec-20	28
Rwanda	Official Gazette nº Special of 21/09/2018 (Law N°48/2018, Law N°49/2018)	Legal Document	13-Dec-20	19
	Official Gazette nº 04 of 23/01/201,221 (LAW N°54bis/2011 OF 14/12/2011 Establishing Rwanda Meteorology Agency (Meteo Rwanda) and Determining Its Mission, Organisation and Functioning)	Legal Document	13-Dec-20	18
	National Policy for Water Resources Management	Official Document	13-Dec-20	34
	Rwanda National Water Resources Master Plan	Official Document	13-Dec-20	35
	LAW N°62/2008 OF 10/09/2008 Putting in Place the Use, Conservation, Protection and Management of Water Resources Regulations	Legal Document	13-Dec-20	20
	Rwanda Meteorology Agency (https://www.meteorwanda.gov.rw/)	Official Website	3-Jan-21	32
	Rwanda Water Portal (https://waterportal.rwb.rw/)	Official Website	3-Jan-21	44
South Africa	South African Weather Service Act 8 Of 2001 and 48 Of 2013	Legal Document	9-Apr-21	22
	Water & Sanitation (https://www.dws.gov.za/)	Official Website	9-Apr-21	7
	National Water Act No 36 of 1998	Legal Document	11-Apr-21	21
	National Water Policy Review (NWPR)	Official Document	11-Apr-21	2
	South African Water Quality Guidelines	Official Document	11-Apr-21	3
	South African Weather Service (https://www.weathersa.co.za/)	Official Website	12-Apr-21	45

Table 1 (continued)

Country	Name of Resource	Resource Type	Date Surveyed	ID
Tanzania	National Water Policy 2002	Legal Document	8-Dec-20	39
	Water Resources Management Act of 2009	Legal Document	8-Dec-20	24
	I-DARE Portal: Tanzania Meteorological Agency	Official Document	8-Dec-20	40
	National Water Sector Development Strategy 2006–2015	Official Document	18-Dec-20	36
	Tanzania Meteorological Authority (http://www.meteo.go.tz/)	Official Website	19-Dec-20	46
Uganda	The Uganda National Meteorological Authority Act, 2012	Legal Document	18-Dec-20	23
	The Data Rescue involving digitization of historical data at Uganda National Meteorological Authority	Official Document	18-Dec-20	51
	Uganda National Meteorological Authority (https://www.unma.go.ug/)	Official Website	19-Dec-20	47
	Ministry of Water and Environment: Republic of Uganda (https://www.mwe.go.ug/)	Official Website	19-Dec-20	38
United States	National Water Information System: Web Interface (https://waterdata.usgs.gov/nwis/qw/)	Official Website	30-Dec-20	48
	NOAA: National Centers for Environmental Information (https://www.ncdc.noaa.gov/)	Official Website	2-Jan-21	42
	Our Strategic Vision for Stewarding the Nation's Climate Data	Official Document	2-Jan-21	41
	CFR Title 40 Chapter I	Legal Document	2-Jan-21	25
	National Water Quality Monitoring Council: Water Quality Portal (https://www.waterqualitydata.us/)	Official Website	3-Jan-21	50
	U.S. Geological Survey (https://www.usgs.gov/)	Official Website	3-Jan-21	49

* *This website is no longer available, so the survey responses were based on an archived version of the site*

tries to find commonalities as well as unique points that contributed to document clarity or straightforwardness from a water resources management standpoint. These elements were defined in a question format for the survey (Table 2). To minimize loss of information that could result due to the use of a binary approach to data translation, the questions were phrased explicitly so that there was very little room left for multiple interpretations. If the answer to the question was implicitly or explicitly presented as "yes" in the official document, the survey value pertaining to that element became 1; if the element could not be found or the answer was "no", the value remained 0 (Fig. 2).

2.3 Data Analysis

For all examined documentation from each country, the summation of unique values for each overarching theme divided by the number of possible elements garnered the representation metric (Fig. 2), essentially an unweighted average by country indicating the proportion of all possible elements represented in at least one document. Since this analysis was

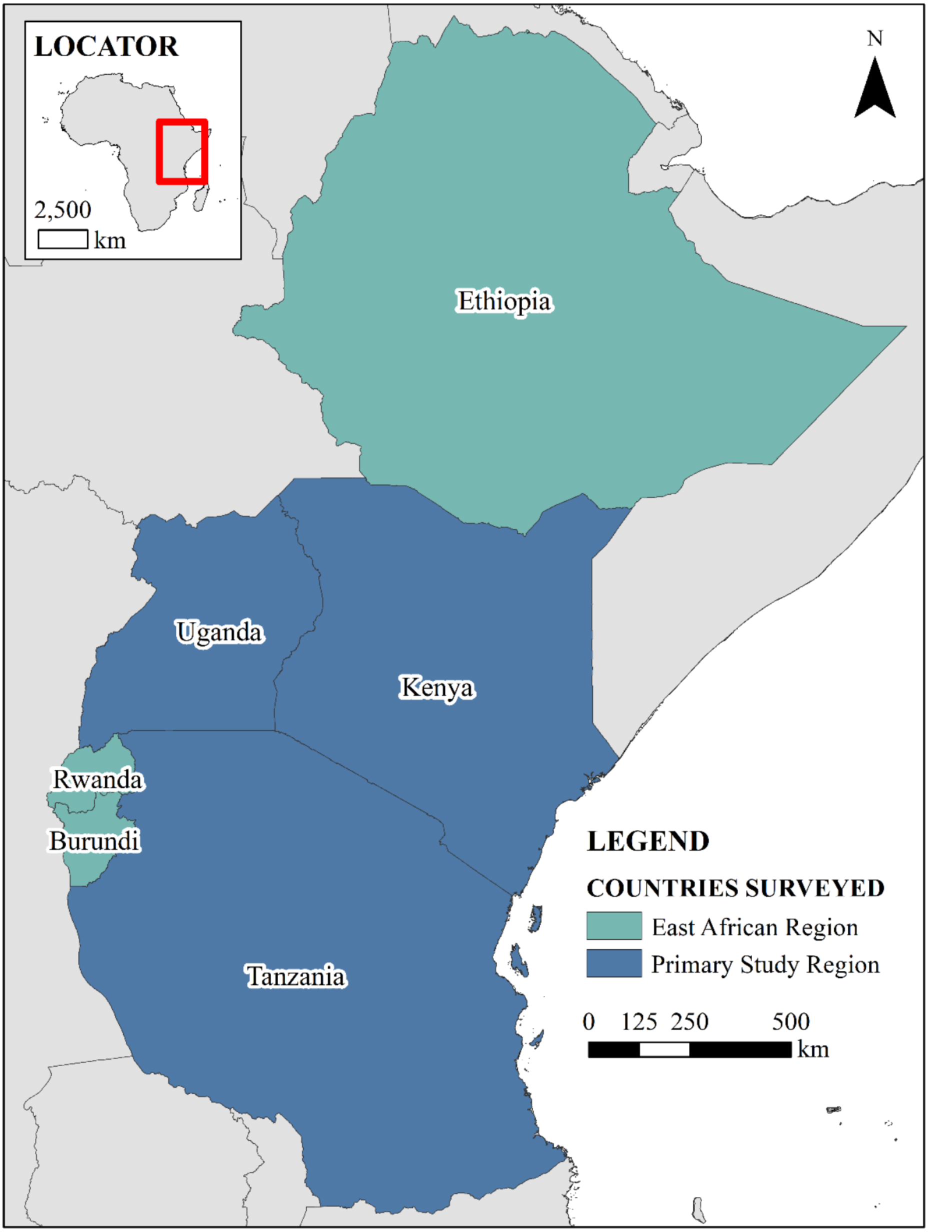

Fig. 1 Map of primary study region and other countries surveyed within East Africa

exploratory, there was no basis for weighting the elements unevenly. If the same element was found in two separate locations, it was recorded but not double counted. This practice was employed to avoid potentially misleading inflation of the respective metric. Metrics ranged from 0 to 1, where 0 indicates that no answers were found for that theme, and a

Table 2 Elements and corresponding questions for each of the three categories

Category	Element	Survey Question*
Data Collection	Format	"Is there a consistency in variables and/or a file format that has been chosen for record keeping?"
	Instrumentation	"Is the instrument or specification of the procedure currently used to take measurements known?"
Data Curation	Digitization	"Is there a plan (or a concern) for the digitization of historical records?"
	Destination	"Is collected data going to be consolidated in a particular location?"
	Service Provision	"Is there a defined strategy to make data publicly accessible online?"
Governance	Responsibility	"Is there a clearly defined entity in charge of collecting, storing, and/or analyzing this type of data?"
	Chain of Command	"Is the structure of entity(ies) in charge of monitoring and recording data measurements defined?"
Performance	Functionality	"Is there evidence of a functioning data access point?"
	Accessibility	"Can data be accessed without special approval?"

** If the answer to the question was implicitly or explicitly presented as "yes", the survey value pertaining to that element became 1; otherwise, the value remained 0.*

perfect score of 1 indicates that all answers and a fully functional and freely accessible data portal were found for that theme.

A graphical comparison was also made of the elements amongst the 30 potential data streams (one for each theme and country) based on whether or not a data stream terminated in a functional, freely accessible maproom or portal for downloading data (i.e., received a value of 1 for Functionality and Accessibility elements). This resulted in a visual representation of what is present or missing in successful data streams and what is present or missing in incomplete data streams. Additionally, the Pearson correlation coefficient (r) for the representation metric and the phi correlation coefficient (φ) for the elements were used to statistically verify the results of the graphical comparison.

3 Results and Discussion

A survey of documentation associated with national management of meteorological, water quantity, and water quality data was conducted to identify predefined elements present for 10 countries. The survey results were summarized using a representation metric to observe overall patterns. A comparison was made between the elements present in the documenta-

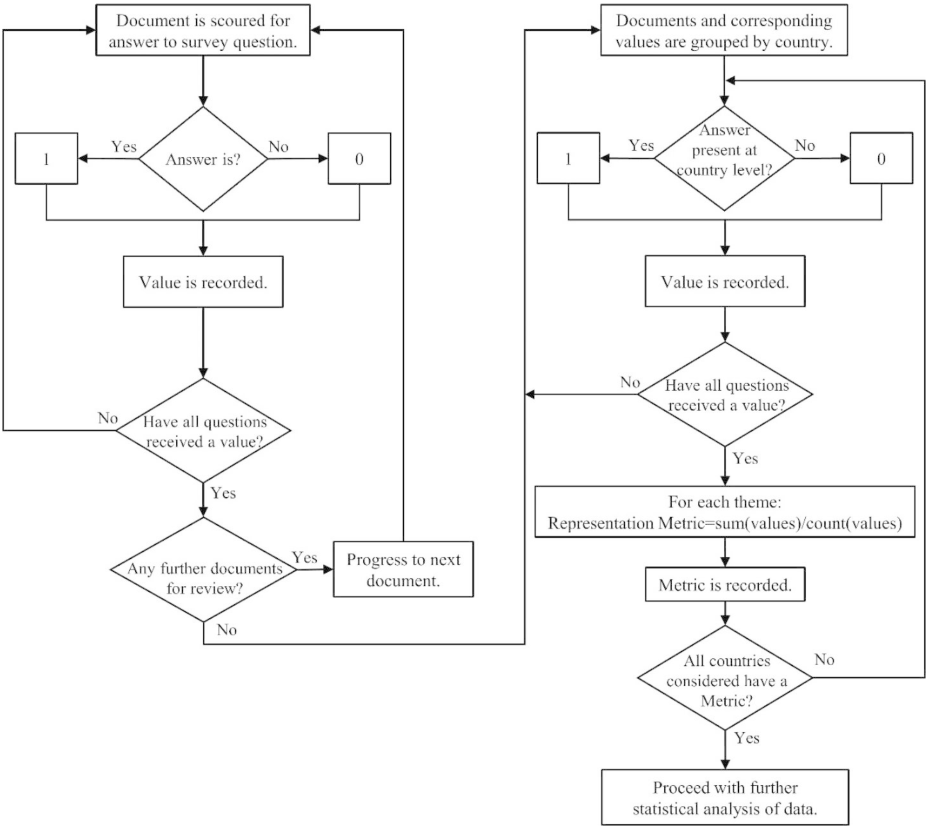

Fig. 2 Flowchart outlining the procedure for surveying documents and calculating the representation metric

tion associated with successful and unsuccessful data streams to investigate what elements or lack of elements are most commonly associated with each type.

3.1 Survey Results

Table 3 shows the representation metrics for each of the potential data streams. Lower values indicate fewer elements present in the documentation for each country and theme, with a maximum value of 1 indicating all elements are present for that data stream. These metrics revealed that, in general, the documentation available for the study region covers fewer elements than the documentation for the external examples (Table 3). However, results showed that there were no perfect performances, suggesting that it is not necessary to have every single element present to achieve a successful data stream.

Figure 3a shows a comparative overview of the presence of elements sought in all data streams separated on a basis of successful performance. A full circle translates to the presence of that element in all data streams in that category and the fractional completion of a circle is a direct reflection of the percentage of data streams for which that element was recorded (e.g., Chain of Command was present for 57% of functional, accessible data streams and 68% of unavailable data streams). Examination of overlap between elements absent in unsuccessful data streams and present in successful data streams suggests that those without a value of 1 for the Service Provision element (i.e., those without a plan or commitment to make data available online), did not result in a functioning and accessible portal. With the exceptions of meteorological data for Tanzania and Rwanda, data streams with the Service Provision element were functional, accessible data streams. The presence of a strictly defined Chain of Command appeared to have little impact on data stream success.

3.2 Analysis of Representation Metrics

To avoid any circularity in the representation metric that could lead to potentially unwarranted inflation, values for the Functionality and Accessibility elements were excluded from calculations related to the comparison of successful and unsuccessful data streams. Successful data streams, marked with an asterisk in Table 3, had an average representation metric of 0.61 compared 0.43 for unsuccessful data streams. The average representation metric for the primary study countries in East Africa (Kenya, Tanzania and Uganda) was 0.51 while the average of all remaining countries was 0.53 (Fig. 3b, left). The corresponding value for all of the surveyed East African countries was 0.42, a lower value than 0.67 for the four external countries (Fig. 3b, right). The discrepancy between successful and unsuccessful data streams suggests that the completeness of documentation available for a data stream is significantly correlated with its functionality and accessibility ($r = 0.7$, $p\text{-value} < 0.0001$). The documentation for all surveyed countries in the East African region exhibited approximately 40% less representation compared with countries that were chosen for their reputations for organized presentation of data and easy to access documentation. The similarity of the average representation metric between countries in the primary study region and the remaining seven countries shows Kenya, Tanzania, and Uganda at a middling position, with an existing foundation as well as opportunities for improved policies, information, and data infrastructure.

Considering all countries, the average representation metrics for meteorological, water quantity, and water quality data were 0.60, 0.50, and 0.47, respectively (Fig. 3c). Although the difference between these averages is not extreme, it highlights numerically that meteorological data streams are generally documented more thoroughly and are more available than water data streams. This is consistent with the experience of searching for information on the three themes, that information on meteorology was generally easier to find than water quality and water quantity. The discrepancy may be in part attributed to the relatively higher effort and cost associated with establishing water quantity and quality monitoring stations. The numbers indicate that the majority of countries considered have the opportunity to develop their water quantity and water quality databases further, an observation that could possibly be extended globally after further research. One convoluting statistic is that there were fewer successful data streams for meteorology than the water quantity and quality themes (Fig. 3d), but this may be attributed to the requisite that successful data streams be both functional and accessible.

Table 3 A summary of comparative documentation and performance for meteorology, water quantity, and water quality data in Burundi (BI), Canada (CA), Ethiopia (ET), Germany (DE), Kenya (KE), Rwanda (RW), South Africa (SA), Tanzania (TZ), Uganda (UG), and the United States (US)

Meteorology Data Streams										
	Country									
	BI	CA*	ET	DE*	KE	RW	SA	TZ	UG	US*
Format	0	1	1	1	0	1	1	1	0	1
Instrumentation	0	0	0	0	1	1	0	0	0	0
Digitization	0	0	0	1	1	0	0	1	1	1
Destination	1	1	0	1	1	1	1	1	1	1
Service Provision	0	0	0	1	0	1	0	1	0	1
Functionality	0	1	0	1	0	0	0	1	0	1
Accessibility	0	1	1	1	0	0	1	0	0	1
Responsibility	1	1	1	1	1	1	1	1	1	1
Chain of Command	0	0	1	1	1	1	1	1	1	1
Representation Metric	0.22	0.56	0.44	0.89	0.56	0.67	0.56	0.78	0.44	0.89
Water Quantity Data Streams										
	Country									
	BI	CA*	ET	DE*	KE	RW*	SA*	TZ	UG*	US*
Format	0	1	0	1	0	1	1	0	1	1
Instrumentation	0	0	0	0	0	1	0	0	0	0
Digitization	0	0	0	0	0	0	0	0	0	0
Destination	0	1	0	1	1	1	0	0	1	1
Service Provision	0	1	0	1	0	1	0	0	1	1
Functionality	0	1	0	1	0	1	1	0	1	1
Accessibility	0	1	0	1	0	1	1	0	1	1
Responsibility	1	1	0	1	1	1	1	1	1	1
Chain of Command	1	0	0	1	1	0	0	1	1	1
Representation Metric	0.22	0.67	0.00	0.78	0.33	0.78	0.44	0.22	0.78	0.78
Water Quality Data Streams										
	Country									
	BI	CA	ET	DE*	KE	RW*	SA*	TZ	UG*	US*
Format	0	1	0	0	0	0	0	0	0	0
Instrumentation	0	0	0	1	1	0	0	0	1	1
Digitization	0	0	0	0	0	0	0	0	0	0
Destination	0	1	0	1	1	1	1	0	1	1
Service Provision	0	0	0	1	0	1	1	0	1	1
Functionality	0	0	0	1	0	1	1	0	1	1
Accessibility	0	0	0	1	0	1	1	0	1	1
Responsibility	0	1	0	1	1	1	1	1	1	1
Chain of Command	1	1	0	0	1	0	1	1	1	1
Representation Metric	0.11	0.44	0.00	0.67	0.44	0.56	0.67	0.22	0.78	0.78
Average Across Themes										
	Country									
	BI	CA	ET	DE	KE	RW	SA	TZ	UG	US
Representation Metric	0.19	0.56	0.15	0.78	0.44	0.67	0.56	0.41	0.67	0.81

**Successful Data Stream: Indicated by the simultaneous presence of Functionality and Accessibility elements*

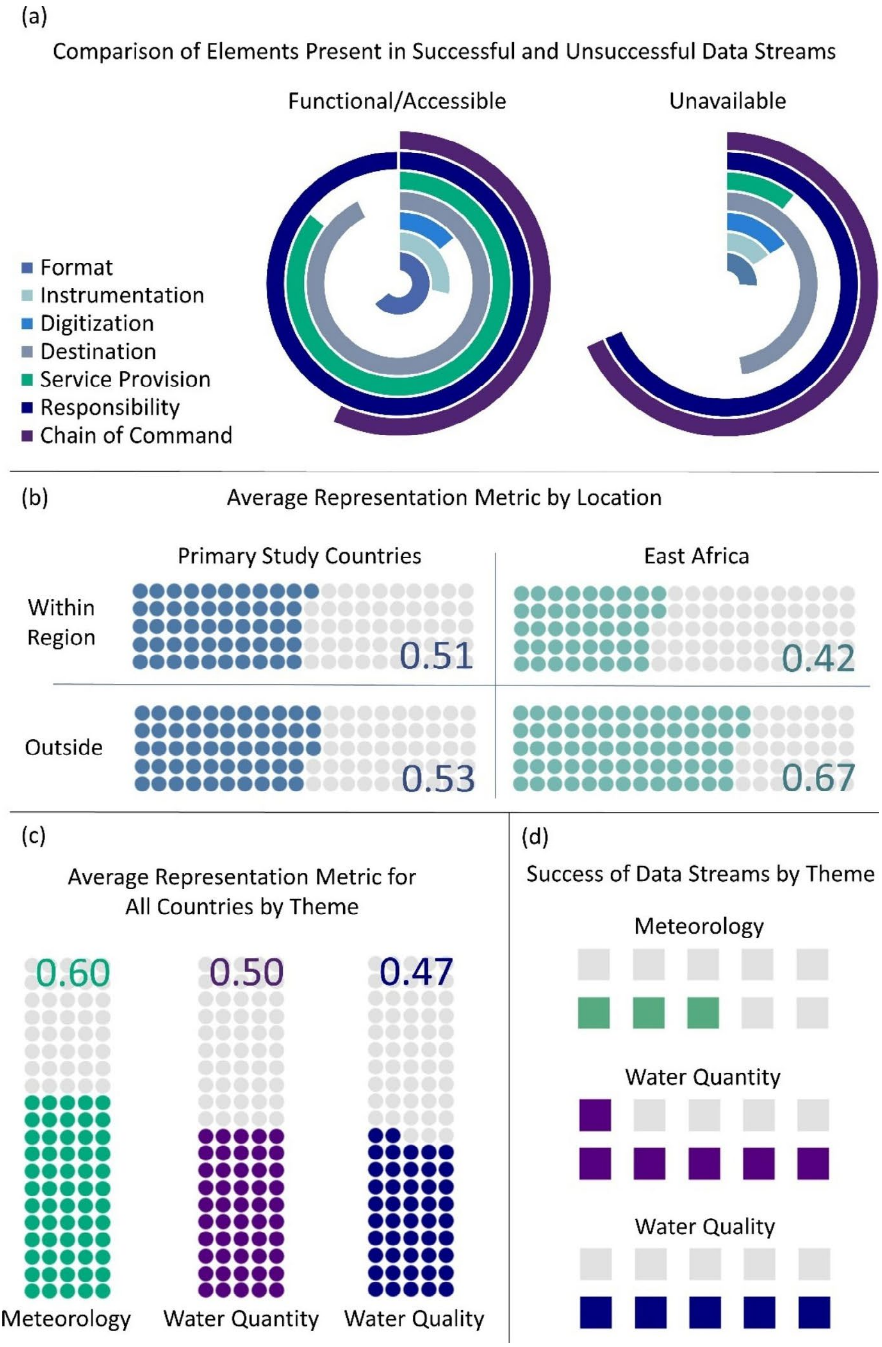

Fig. 3 Graphic illustrating (a) a comparison of documentation elements present in functional, accessible data streams versus data streams that did not meet those requirements, (b) a comparison of the average representation metric within and outside of regions consisting of the primary study countries and the East African countries, (c) a comparison of average representation metrics by theme, and (d) a series of waffle charts demonstrating actual versus potential data stream success based on theme

3.3 Data Collection Elements

The Format element was introduced to observe different modes of unifying records for more efficient management of datasets. The majority of format element instances took the form of a defined list of variables. In most cases with an accessible database, a definitive list was not present, but the data entries for variables measured and housed in the database was formatted consistently such that an implied list of variables could have been derived. Some examples of the defined variables list include: Die deutschen Klimabeobachtungssysteme (2017), which detailed the variables considered by the Global Climate Observing System (GCOS) crossing the themes of meteorology and water quantity; and, Ethiopia and Canada meteorology websites which had a page of dataset variables and their definitions. For others, the element consisted of a simple list. Explicit lists of variables were typically only found regarding meteorology. Details extending to the format in which the data was to be stored were very rare. Of the functioning, accessible data streams, approximately 2/3 had an associated format element. The Format element, while not strictly essential to the creation of a functioning data portal is still related; for example, one of the expectations Rwanda has of their planned database is that it "Standardiz[es] the data format for meteorological and hydrological data" (MINIRENA-RNRA 2015). The development of an organized, functional database and the determination of which data variables it stores and how these records are entered and displayed are logical accompaniments to one another. An explicit list defining the meaning of these variables is a valuable resource to potential users, preventing confusion over the meanings of specific measurements.

The Instrumentation element was designed to be a check on data measurement quality and consistency as well as the knowledge level of the providers regarding these properties of their data. There were few examples of this element; of those existing, most only had basic information regarding the type of instrument used to gather data. Some made indirect reference to the standards set out by the World Meteorological Organization (WMO) (KMD 2021; DWD 2021). Knowledge of the instruments used to collect data did not appear to have a strong relationship to data stream success. It is possible that since databases are typically filled internally, service providers see no need to divulge this level of detail as, presumably, they are confident with their own measurement systems. However, the existence of quality standards for information allowed into a database, regardless of whether it is internally or externally sourced, is critical when developing secondary products and interpreting results of analyses. Without an understanding of the limitations of the data, there is a risk of being misled about its quality because it was routed through a trusted database.

3.4 Data Curation Elements

Digitization was a concern expressed mostly for meteorological data records. Generally, the presence of Digitization in official government sources was limited to a single statement rather than a full report. In the primary study countries these preservation projects were primarily driven and sponsored by international organizations. For example, an I-DARE document was included for Tanzania because it had an official seal in the header, but the project was conducted in partnership with the UK Met Office and funded by the DFID (MWTC undated). Some documents which covered similar projects were not included because they were not strictly from official government sources, e.g., USAID project report, Tanzania

(Tetra Tech, Inc. 2016), NOAA African Upper Air Data Rescue, Kenya (WMO 2008). Historical data is an important part of climate and water research, and in many cases, the preservation of records is time-sensitive due to the deterioration of paper records. Provided that records remain viable, Digitization is an element worthy of the allocation of sufficient resources to transform them into usable formats.

The Destination element was contrived to isolate a key conceptual piece required in progress toward a functioning data access point. Many data streams either had a designated database or intended to develop a database for the storage of their climate and water data, with presence of the element in 100% of successful data streams and only 52% of unsuccessful data streams. In some cases, Destination involved the consolidation of several pre-existing databases into a single main source. This is a notable concern for governments for which the responsibility of collecting a single theme of data is or has been split amongst multiple entities in the past (e.g., MINIRENA-RNRA 2015, USGS et al. 2021). Unifying similar databases and aggregating data makes the most of existing resources, ensuring that available information does not get lost or overlooked.

A commitment to making data readily and freely available to potential researchers is another notable piece when targeting the realization of a successful data stream. This foundational concept was captured by the Service Provision element. There were only two data streams, Tanzanian and Rwandan meteorology, for which the presence of Service Provision did not result in a working, accessible portal, indicating the importance of the element and reflecting its strong correlation ($\varphi = 0.75$) with the desired result of accessible data.

3.5 Governance Elements

The Responsibility element covers a rudimentary aspect of a successful data collection stream, that there must be an agency, ministry, or similar organization charged with the collection of data. It follows, but is not always essential, that this same entity be tasked with the management and dissemination of the data once it is collected. All successful data streams, versus 62% of unsuccessful data streams, had a document which listed what entity was responsible for collecting a particular data type. For many countries, there is some overlap in the agencies responsible for collecting water quantity and quality data, potentially to promote cost efficiency or to form a bridge between organizations as their roles change or one is phased out. The lack of detail does, however, add some confusion as to who is ultimately responsible for data dissemination and which portal is the most direct source. In many cases, responsibilities such as data collection and development of long-term water resources management plans have been decentralized (Government of the Federal Republic of Germany 2018; WRA 2019; MWLD 2002; Government of the United Republic of Tanzania 2009; MoWI 2008).

Chain of Command was an element designed to further assess the management and organizational styles of different nations. In the past decade, the primary East African countries have been establishing agencies to officially assume the task of monitoring meteorological and water data and providing public services (Government of the Federal Republic of Germany 2020; KMD 2019; Government of the Republic of Uganda 2012). Often, some form of the entity was already in existence but underwent a renaming and/or restructuring with the addition of a new act or policy. For example, in 2019, the then Tanzania Meteorological Agency was renamed and transformed into Tanzania Meteorological Authority. In these instances, the upper-level structures of the entities and the general duties of its components were well documented and were intended to accommodate and integrate rather than preclude others from collecting similar data provided that they had permission to do so. Based on the comparison shown in Fig. 3a, supported by the calculated correlation coefficient (φ = -0.12), there is not a strong positive correlation between Chain of Command and a functional, accessible data portal. However, this element was one of the more thoroughly covered elements in data streams lacking either functionality or accessibility. While this does not imply a strong negative correlation, it could indicate that the definition of a strict, elaborately defined bureaucratic structure could potentially hinder data-driven decision-making and management.

3.6 Performance Elements

The Functionality and Accessibility elements were designed to reflect the user experience as a researcher unaffiliated with the entities in charge of collecting data and the ultimate success of the data stream. The easiest databases to categorize were those which, after an experimental attempt to procure data, produced a file download in an easily legible form (comma separated values in all such cases). A variation on this were those that required that a user log in to access the database but granted login credentials without requiring a potential user to submit special documents (i.e., the WIS databases for Uganda [MWE 2021]) ending in the same result. The data collection entities which presented a data request form or request via email (e.g., Kenya [KMD 2021], and South Africa [DWS 2021]) either provided no response to the request or asked for a formal letter of introduction from a research institution. One website invited users to request access while requiring a similar introduction document as well as a copy of an official ID or passport (Meteo Rwanda 2021). There are indications that countries in the primary study region are working on improving online access to data. For the Tanzania Meteorological Authority (TMA) and the Uganda National Meteorological Authority (UNMA), this has taken the form of an interactive maproom. However, attempts to open and utilize these maprooms were unsuccessful. Websites which were not loading after multiple attempts on different networks and browsers, required user review and approval for access, or did not produce a download when the download link was activated were considered non-functional and inaccessible.

4 Additional Discussion and Recommendations

This work was based on documentation that was available, openly accessible, and as current as possible; it is acknowledged that other information may exist beyond these documents. However, experience with the portals provided a form of confirmation that the documentation considered was fairly consistent with what was available. Many documents and resources were available in English, which greatly enhanced their accessibility. In the instances where official translations were not available, alternative translation tools such as those featured in Microsoft Edge and Google Translate were used with enough efficacy that the document substance was not greatly diminished. Documents were only analyzed at the national level for this study, and it remains to be seen whether more specifically targeted searches would answer remaining questions.

The representation metrics, as developed for this study, are a summarization of massive amounts of qualitative data. In this regard, a binary or forced choice survey could potentially lead to a loss in specificity. However, it is also one of the most straightforward methods for collecting data (Dolnicar et al. 2011). To build upon this study, it would be prudent to explore options for creating levels beyond a binary scale for survey responses to see if more trends and linkages emerge. The addition of more countries would provide further evidence that observed correlations transcend a particular region. In this study, a broad focus was maintained on meteorological, water quantity, and water quality data. While beyond the scope of this study, a breakdown of specific data types, time scales, and spatial densities of data available at the endpoint of successful data streams could also be very useful in evaluating the overall quality of accessible data and providing insights on strengths and deficiencies in existing data infrastructure at the country level.

4.1 Implications

Research is an open-ended domain, with multiple facets and an ever-changing landscape. Water resources management is currently being pitted against various dynamics in both quality and quantity fronts in a changing climate. This calls for continuous concerted research endeavors with access to authenticated hydrometeorological data being a key pillar for successful generation of requisite analytics to inform decision-making processes.

Any dataset remains of little use unless translated into applicable information through an analytical process. The more data is freely made available, the wider the spectrum of analytics that can be generated by different professionals and practitioners at minimal cost. This increases the scope and possibilities of research leading to availability of an array of information for use by decision makers. It also facilitates efforts to reference and archive related research and accompanying analytics, thereby reducing repetition of efforts and eliminating redundancies. Likewise, any decision in water resources management not backed by requisite information or analytics has a high chance of failing. In this case, making more information available and accessible to decision makers in user-friendly formats results in wiser decisions for economic development and the benefit of future generations.

4.2 Challenges and Opportunities for Improving Data Availability, Access, and Use

As governments focus on delivering the needed development services and infrastructure, data collection, management, and dissemination do not receive the deserved attention. Due to fiscal and structural limitations, data collection is often not regular and systematic, and data management procedures are not streamlined. Further, large spatial and historical gaps exist in the data, limiting abilities to assess trends and make forecasts. Variation in mandates causes data to be scattered across entities in databases that are not necessarily linked. Consequently, the data available to the public is time-consuming to find; navigating and translating the high-level aggregation and analysis of relationships and interlinkages is often difficult (e.g., land use impacts on water quality and quantity at watershed level). Technological advancements enable automation of data collection, reducing the cost and increasing accuracy. Along with this advancement are diversified uses for the data products. To maximize the value of collected datasets, they can be packaged in forms relevant and accessible to a variety of stakeholders.

Good quality data is not only essential for decision-making and management of resources but a critical necessity for research associated with water resources. For example, hydrologic and water quality models are tools widely used for research aspects such as evaluation of land use, climate, and conservation practice impacts on watershed responses. Availability of quality data for precipitation and temperature, observed hydrological metrics (e.g., streamflow or discharge), water quality metrics (e.g., sediment loads, nitrogen and phosphorus concentrations) is crucial for model calibration and validation in order to develop reliable model outputs that provide valuable information for making sound management decisions and policies (Holling 1978; Kirchner 2006). Findings from this study provide optimism grounded in the possibility of better databases being made available in the study region and beyond as a result of the study recommendations.

The lack of funding in the primary study countries is an additional challenge faced by departments, ministries, and agencies tasked with providing climate and water data. This issue has been acknowledged in internal assessment reports (e.g., MoWI 2008, WRA 2019, MWE-DWRM undated). Insufficient budgets and resources can lead to the instatement of other means of self-sufficiency such as charges placed on data products or specialized analyses (Oui 2021). A frequently cited justification for an open data policy is that it facilitates economic growth, development, and innovation (Borgesius et al. 2015). The commodification of data has pros and cons in that it allows the market to drive what products are created (Oui 2021) but also creates a situation where research can be made cost prohibitive, particularly if charging for raw data. Such a setting defeats the primary purpose of establishing hydrometeorological monitoring systems.

5 Conclusions

The discrepancy between the average representation metrics of available and unavailable data streams indicates that the presence of more elements is correlated with data stream success. Overall trends in documentation suggest that processes surrounding water quantity and quality data are not as well documented as meteorology data. While relevant documentation for the primary study region was, when averaged as a representation metric, comparable to that of the other seven countries, it generally presented fewer elements than the documentation for the comparison region comprised of select countries external to East Africa, suggesting need for improvement. The goal of this analysis is not to suggest that current strategies in the primary countries in East Africa are ineffective but rather to provide concrete opportunities to build upon their existing systems using the experience of each other and others as an example that can potentially be refined even further before application. The main elements associated with a freely accessible data stream were Format, Destination, Responsibility, and Service Provision, from which it can be inferred that consistency in data variables, a defined database where the data will be stored, a clear idea of what organization will take the lead on data collection and dissemination, and above all a documented commitment to providing data access are the keys to success.

Authors' contributions V. M. Garibay and M. W. Gitau designed the framework for the analysis. V. M. Garibay conducted the analysis. V. M. Garibay and M. W. Gitau wrote the manuscript in consultation with V. Kongo, J. Kisekka, and D. Moriasi, who provided valuable insights from their international, interdisciplinary perspectives. All authors reviewed the results and contributed to the final version of the manuscript.

Funding This work was funded in part by the LASER-PULSE program through USAID's Innovation, Technology, and Research Hub.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to Publish Not applicable.

Conflict of interest The authors declare no potential conflicts of interest.

6 References

Aman M, Abdulkadir SJ, Aziz IA, Alhussian H, Ullah I (2021) Multimedia Tools and Applications 80(8):12469. https://doi.org/10.1007/s11042-020-10215-x. KP-Rank: A semantic-based unsupervised approach for keyphrase extraction from text data

Borgesius FZ, Gray J, Van Eechoud M (2015) Open Data, Privacy, and Fair Information Principles. Berkeley Technol Law J 30(3):2073–2131. https://doi.org/10.15779/Z389S18

Bui DDA, Del Fiol G, Hurdle JF, Jonnalagadda S (2016) Extractive text summarization system to aid data extraction from full text in systematic review development. J Biomed Inform 64:265–272. https://doi.org/10.1016/j.jbi.2016.10.014

Dolnicar S, Grün B, Leisch F (2011) Quick, Simple and Reliable: Forced Binary Survey Questions. Int J of Market Res 53(2):231–252. https://doi.org/10.2501/IJMR-53-2-231-252

DWD (2017) Die deutschen Klimabeobachtungssysteme. https://gcos.dwd.de/DWD-GCOS/DE/publikationen/inventarbericht/gcos_inventarbericht_2013.pdf

DWD (2021) Deutscher Wetterdienst: Wetter und Klima aus einer Hand. https://www.dwd.de/

DWS (2021) Water & Sanitation. https://www.dws.gov.za/

Gebrechorkos SH, Hülsmann S, Bernhofer C (2018) Evaluation of multiple climate data sources for managing environmental resources in East Africa. Hydrol Earth Syst Sci 22:4547–4564. https://doi.org/10.5194/hess-22-4547-2018

Garibay VM, Gitau MW, Kiggundu N, Moriasi D, Mishili F (2021) Water Resour Manage 35(5):1587–1602. https://doi.org/10.1007/s11269-021-02804-8. evaluation of reanalysis precipitation data and potential bias correction methods for use in data-scarce areas

Gitau MW, Gburek WJ, Jarrett AR (2005) A tool for estimating best management practice effectiveness for phosphorus pollution control. J Soil Water Conserv 60(1):1–10

Government of the Federal Republic of Germany (2018) Wastewater Charges Act of 2005, Federal Law Gazette I p. 114. Amended 2018, Federal Law Gazette I p. 1327. https://www.gesetze-im-internet.de/abwag/AbwAG.pdf

Government of the Federal Republic of Germany (2020) Water Resources Act of 2009, Federal Law Gazette I p. 2585. Amended 2020, Federal Law Gazette I, p. 1408. http://www.gesetze-im-internet.de/whg_2009/WHG.pdf

Government of the Republic of Uganda (2012) The Uganda National Meteorological Authority Act, 2012. http://faolex.fao.org/docs/pdf/uga158640.pdf

Government of the United Republic of Tanzania (2009) Water Resources Management Act, 2009. http://extwprlegs1.fao.org/docs/pdf/tan96340.pdf

Holling CS (ed) (1978) Adaptive environmental assessment and management. Wiley, London

Kirchner JW (2006) Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour Res 42:W03S04. doi:https://doi.org/10.1029/2005WR004362

KMD (2019) Meteorology Policy. http://www.environment.go.ke/wp-content/uploads/2019/05/DRAFT-METEOROLOGY-POLICY-2019.pdf

KMD (2021) Kenya Meteorological Department. Republic of Kenya. https://www.meteo.go.ke/

Lakew HB, Moges SA, Anagnostou EN, Nikolopoulos EI, Asfaw DH (2020) Evaluation of Global Water Resources Reanalysis Runoff Products for Local Water Resources Applications: Case Study-Upper Blue Nile Basin of Ethiopia. Water Resour Manage 34(7):2157–2177. https://doi.org/10.1007/s11269-019-2190-y

Mason S, Kruczkiewicz A, Ceccato P, Crawford A(2015) Accessing and Using Climate Data and Information in Fragile, Data-Poor States. [Report] IISD, Winnipeg. https://www.iisd.org/system/files/publications/accessing-climate-data-information-fragile-data-poor-states.pdf

Meteo Rwanda (2021) Rwanda Meteorology Agency. https://www.meteorwanda.gov.rw/

MINIRENA-RNRA (2015) Rwanda National Water Resources Master Plan. https://www.rwb.rw/fileadmin/user_upload/RWRB/Publications/Policies/Rwanda_National_Water_Resources_Master_Plan.pdf

Mogaka H, Gichere S, Davis R, Hirji R(2006) Climate variability and water resources degradation in Kenya: improving water resources development and management. World Bank working paper No. 69. Washington, D.C. https://doi.org/10.1596/978-0-8213-6517-5

MoWI (2008) National Water Sector Development Strategy: 2006–2015. http://extwprlegs1.fao.org/docs/pdf/tan169532.pdf

MWE (2021) Ministry of Water and Environment: Republic of Uganda. https://www.mwe.go.ug/

MWE-DWRM (undated) Management Strategy and Action Plan: Upper Nile Water Management Zone. https://www.mwe.go.ug/sites/default/files/library/UNWMZ%20Strategy%20%26%20Action%20Plan.pdf

MWLD (2002) National Water Policy. http://www.tzonline.org/pdf/waterpolicy20021.pdf

MWTC (undated) The United Republic of Tanzania Ministry of Works, Transport and Communication: Tanzania Meteorological Agency. https://www.idare-portal.org/sites/default/files/I-DARE_PORTAL_TANZANIA.pdf

Näschen K, Diekkrüger B, Leemhuis C, Steinbach S, Seregina LS, Thonfeld F, Van der Linden R (2018) Hydrological Modeling in Data-Scarce Catchments: The Kilombero Floodplain in Tanzania. Water 10:599. https://doi.org/10.3390/w10050599

Oui J (2021) Commodifying a "Good" Weather Data: Commercial Meteorology, Low-cost Stations, and the Global Scientific Infrastructure. Sci Technol Hum Values: Advance Online Publication. https://doi.org/10.1177/01622439211995889

Overpeck JT, Meehl GA, Bony S, Easterling DR (2011) Climate Data Challenges in the 21st Century. Science 331:700–702. https://doi.org/10.1126/science.1197869

Tetra Tech, Inc (2016) PREPARED: Tanzania Meteorological Agency Climate Data Rescue Pilot Project Report. https://www.climatelinks.org/sites/default/files/asset/document/2016_USAID-PREPARED_Tanzania-Meteorological-Agency-Climate-Rescue-Pilot-Project.pdf

Tian F, Wu B, Zeng H, Ahmed S, Yan N, White I, Zhang M, Stein A (2020) Identifying the Links Among Poverty, Hydroenergy and Water Use Using Data Mining Methods. Water Resour Manage 34:1725–1741. https://doi.org/10.1007/s11269-020-02524-5

Turksen IB, Celikyilmaz A(2010) Information Extraction from Text – Dealing with Imprecise Data in Azar AT (ed.) Fuzzy Systems. 147–168. https://doi.org/10.5772/7223

Wang WM, Li Z, Wang JW, Zheng ZH (2017) How Far We Can Go with Extractive Text Summarization? Heuristic Methods to Obtain near Upper Bounds. Expert Syst Appl 90:439–463. https://doi.org/10.1016/j.eswa.2017.08.040

WMO (2008) Report of the meeting of the CCl Expert Team on the Rescue, Preservation and Digitization of Climate Records. https://library.wmo.int/doc_num.php?explnum_id =9411

The World Bank (2019) Open Data Essentials. http://opendatatoolkit.worldbank.org/en/essentials.html

WRA (2019) Water Resources Authority Strategic Plan: 2018–2022. Republic of Kenya. http://extwprlegs1.fao.org/docs/pdf/ken207659.pdf

Zaeem RN, German RL, Barber KS (2018) ACM Trans Internet Technol 18(4):53:1–5318. https://doi.org/10.1145/3127519. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining