DMP du projet "Hydrologic Extremes at the Global Scale (HEGS)"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "Horizon 2020 FAIR DMP (anglais)" fourni par Commission européenne.

Plan Details

Plan title	DMP du projet "Hydrologic Extremes at the Global Scale (HEGS)"
Language	fra
Creation date	2019-10-30
Last modification date	2021-06-28
Identifier	HEGS

Project Details

Project title

Hydrologic Extremes at the Global Scale (HEGS)

Abstract

Hydrologic extremes (floods and intense precipitations) are among Earth's most common natural hazards and cause considerable loss of life and economic damage. Despite this, some of their key characteristics are still poorly understood at the global scale. The IPCC thus reports "a lack of evidence and thus low confidence regarding the sign of trend in the magnitude and/or frequency of floods on a global scale". More generally, the space-time variability of hydrologic extremes is yet to be thoroughly described at the global scale. As a striking illustration, the recent initiative "23 unsolved problems in Hydrology that would revolutionise research in the 21st century" of the International Association of Hydrological Sciences includes questions such as: are the characteristics of extreme hydrologic events changing and if so why? How do extremes around the world teleconnect with each other and with other factors? Why do extreme-rich/poor periods exist?

It is vital to fill these knowledge gaps to inform design, safety and financial procedures and to improve hazard preparedness and response. The project's ambition is hence to better understand the global space-time variability of hydrologic extremes, using a three-pillar research strategy based on methodological innovation, extensive data analysis and proof-of-concept case studies. The specific objectives are to:
1. Develop a statistical framework to describe the global-scale variability of extremes in relation to climate;
2. Analyse global precipitation/streamflow datasets with the aim of quantifying teleconnections, spatial clustering, trends and extreme-rich/poor periods, along with their climate drivers;
3. Explore practical applications such as global early warning systems allowing international disaster response organisations to trigger early actions.

Successful completion of the project will deliver new tools to analyse extremes at the global scale and will hence contribute to more efficient risk management.

Funding

- European Commission-H2020-MSCA-IF-2018 : 835496

Research outputs :

1. Probabilistic modeling of data varying in Space, Time or other Dimensions (Logiciel)

Contributors

Name	Affiliation	Roles
Benjamin Renard - https://orcid.org/0000-0001-8447-5430		• Coordinateur du projet • Personne contact pour les données • Responsable du plan

DMP du projet "Hydrologic Extremes at the Global Scale (HEGS)"

---

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

Collected data: existing hydro-meteorological datasets
The analyses planned in this project are based on existing datasets of hydro-meteorological variables such as precipitation, streamflow, atmospheric pressure, temperature, etc. The table below summarizes the properties of the main global-scale datasets we intend to use.

Dataset	Variable	Space/time resolution	Extent	Period
Hadex2	Precipitation	Station data, monthly extreme indices	Global	Station-specific
Hadex3	Precipitation	Station data, monthly extreme indices	Global	Station-specific
ADHI	Streamflow	Station data, Annual indices	Africa	Station-specific
GSIM	Streamflow	Station data, monthly indices	Global	Station-specific
GRDC	Streamflow	Station data, daily series	Global	Station-specific
NCEP/NCAR	Atmospheric	Gridded data, daily	Global	1948-present
ERA40	Atmospheric	Gridded data, daily	Global	1957-2002
20CR	Atmospheric	Gridded data, daily	Global	1836-2015
ERA20C	Atmospheric	Gridded data, daily	Global	1900-2002
ERSST	Sea Surface Temperatures	Gridded data, monthly	Global	1854-present
Kaplan	Sea Surface Temperatures	Gridded data, monthly	Global	1856-present

In addition to these global-scale datasets that are closely related to the main objectives of the project, we may also use national-scale (most likely Australia and France) datasets containing precipitation/streamflow times series for testing purposes. This DMP will be updated with a more precise description of these datasets once decisions have been made, but in any case we will restrict ourselves to openly-accessible datasets.

Generated data: codes and software
One of the main objectives of the project is to develop a methodological framework for predicting hydrologic extremes using large-scale climate information. This framework will be implemented into a software tool comprising two main components:

1. a computational engine in the form of a FORTRAN-based executable file, piloted by text configuration files;
2. a user interface to pilot this computational engine, based on a set of user-friendly functions coded in the language R.

Both components will be released as open-source software for further use by the hydro-climatic community.

Generated data: Results
This project is not expected to generate new observation or reanalysis datasets per se; however, the results of the case studies may be released in various forms and may be reused by others. While the precise content of these 'results' datasets is unclear at this stage, the following types of datasets may be used:

1. Raw values files (typically text files);
2. Media (images, sounds and movies);
3. Interactively-browsable results (typically interactive maps displayed in the internet browser).

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **Outline the discoverability of data (metadata provision)**
- **Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

1. **Existing datasets** are well-known to the hydroclimate community, are documented in the scientific literature and are referenced in climate data guides. They follow metadata standards used by the hydroclimatic community (typically CF metadata).
2. **Project-generated codes and software** will be deposited to and referenced (with a unique identifier) by the organization for the protection of authors and publishers of digital creations. They will be released through GitHub (which includes version control tools) and Zenodo. Some R packages may also be deposited as packages on the Comprehensive R Archive Network, which involves thorough checks in terms of metadata completeness, documentation, operating system portability, etc.
3. This DMP will be updated later during the project to describe in more details the metadata and documentation associated with the **'results' datasets**.

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

1. **Existing datasets** are all openly available on the internet (links are provided in the table of section 1). Methods and software tools to access, read and manipulate them are numerous, openly accessible and well-documented. In this project we mostly use R packages to handle files in the netcdf format.
2. **Project-generated codes and software** will be made publicly available through GitHub and the CRAN repository.
3. This DMP will be updated later during the project to provide more specific access details for the **'results' datasets**. We aim at making them all openly available, and we don't foresee access restrictions or specific software needs (beyond standards tools such as media player, web browser etc.). Distribution means include websites (e.g. the project's blog), data open-access repositories such as Zenodo, data papers, etc.

2.3 Making data interoperable:

- **Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.**
- **Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?**

1. **Existing datasets** are based on standard formats used by the hydroclimatic community (typically netcdf format).
2. **Project-generated codes and software** will be usable on most operating systems. Compiled code will be compiled for both Linux and Windows, and since the code will be open, others may recompile it for other OS. R packages are portable on multiple OS.
3. This DMP will be updated later during the project to provide more details for the **'results' datasets**. Overall, we will aim at using either simple formats that can be reused by any system (e.g. column-structured text files, json, etc.) or widely-used formats supported by most systems (e.g. netcdf, standard media formats, etc.).

2.4 Increase data re-use (through clarifying licenses):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

1. **Existing datasets** can be re-used with no particular restriction or duration limitation (at the time of writing of this DMP). The licenses and access conditions are specific to each data product (for details, see links in the table of section 1).
2. **Project-generated codes and software** will be released with a GPL-3 open-source license. We will make every effort to develop a thorough documentation, which is a key component for code/software re-usability.
3. Specific licenses and access conditions for **'results' datasets** will be updated later, but again, our objective is to use licenses allowing a non-restrictive use of data.

---

3. Allocation of resources

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

Benjamin Renard (PI) is responsible for data management in the project. Since this DMP is based on existing institutional repositories and portals, we do not foresee any direct cost. Indirect costs such as the time spent for making data FAIR or the cost of maintaining institutional repositories will be covered as part of the project's fixed-amount institutional costs.

---

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

1. **Existing datasets** can be recovered at any time from the internet. In addition, they are copied locally on two hard drives.
2. All **project-generated codes** are stored on GitHub repositories, which ensures safe and backed-up storage, remote access, version control, etc.
3. The main **'results' datasets** will be stored on institutional file-hosting systems at the University of Adelaide and/or Irstea during the projects' lifetime. The DMP will be updated with longer-term solutions once the precise content of these 'results' datasets is known and corresponding storage decisions have been made.

---

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

None

---

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

This DMP has been prepared with the OPIDoR portal developed by Inist-CNRS.