1. Administrative information

1.1 Your name (first name and surname)?
Jacob Qvam Skavang

1.2 Master Thesis Project Title
Assessing the Shyft Modelling Framework in Nepal: Impact of Snow Routines and Terrain Representation on Simulated Water Balance Components

1.3 Supervisor(s)
Enter full name and affiliation of your supervisors
Olga Silantyeva, UiO; Lena Merete Tallaksen, UiO; Kristoffer Aalstad, UiO

1.4 Project description
Give a short summary of the project.
In addition, where is the research being carried out, and what is under study? Is the study individually based, part of a larger project, or being carried out in cooperation with an institution(s), e.g. a museum(s), a state/regional authority, or community group(s), etc? Give as much information about the project as possible.

The region encompassing the Himalayan and Tibetan Plateau is of immense significance as it provides water resources for millions of people residing in the surrounding areas (Bookhagen & Burbank, 2010). The water from this region provides drinking water, supports agricultural activities, is used for generating hydro power, and catering to other agro-economic requirements (Ménégoz et al., 2013). The area holds tremendous potential for developing hydro power and thus making the transition towards a greener economy more feasible.

The Budhi Gandaki catchment, located in Nepal's Gorkha district, receives a high mean annual precipitation of 1495 mm and has an extremely diverse topography, with elevations varying from 479 to 8163 meters above sea level (Devote et al. 2017). Its unique characteristics make it an attractive location for hydro power development, with an installed capacity of 1200 megawatts (MW) and an average energy generation of 3383 gigawatt-hours (GWh). However, the hydro power potential is subject to the prevailing climatic conditions in precipitation, evaporation, temperature, and snow/ice within the catchment, as noted in Edenhofer et al 2011. Climate change can have serious implications to the hydro power production (Dandekhya et al. 2017).
Changing rainfall pattern and increased temperatures will affect power generation. The retreat of glaciers, expansion of glacial lakes, and alterations in the seasonality and intensity of precipitation are some of the factors that will impact power generation in the future (Dandekhya et al. 2017). In addition, the region is also susceptible to Glacial Lake Outburst Floods (GLOFs), which can cause devastating floods downstream (Dandekhya et al. 2017).

The complex hydrology of the High Mountain Asia poses challenges to hydrological modelling due to highly variable distribution of hydro-meteorological variables, steep elevation gradients and pronounced seasonality by the Indian Monsoon, as well as differing moisture regimes between the Tibetan Plateau and the Indian Ocean region (Bhattarai et al., 2020; Fan et al., 2019). Additionally, climate change can likely alter the regional water balance components (Cachet al., 2011). Accurate measurement of meteorological variables is challenging due to the region's heterogeneous topography (Pelliciotti et al. 2012). It is nearly impossible to make observations at all elevations, especially for snow. Insufficient data can lead to errors in discharge predictions, making satellite observations and reanalysis data sets vital for the decision-making processes. Furthermore, the input data and choice of models can have a considerable impact the accuracy of the outcome (Kaufeldt et al., 2016).
In the past, several hydrological models have been developed for various applications. Some of these models include SRM (Martinec et al., 2021), HEC-HMS (Halwatura and Nanjim 2013) and Shyft (Burkhart et al. 2021). Unlike other many other models, Shyft offers the advantage of integrating various well-known hydrological routines (each defined for processes such as evapotranspiration, snow accumulation and melt) into multiple models. This allows for testing multiple different scientific hypotheses using only this framework. Furthermore, it is important that a model can accurately represent terrain features since the terrain in High Mountain Asia is complex. Most distributed models discrete the catchment and derive terrain features from Digital

Elevation Models (DEMs) (Bhattarai et al. 2020). Shyft is able to use several different catchment discretisation methods, such as regular grid and Triangular Irregular Networks (TINs). TIN-based models have a better accuracy in representing complex terrain than grids, while also being computationally efficient. Few studies have used TINs in Himalayan catchments, but Bhattarai et al. (2020) has shown that TINs can improve simulation results compared with grids. Shyft is therefore considered a particularly valuable tool for its ability to evaluate different models, and to consider terrain heterogeneity and generate distributed output variables. Shyft also aims to become a FAIR framework for learning and research activities promoting open-science efforts Wilkinson (2016), which makes it ideal for scientific purposes.

Aim and objectives

The primary aim of this study to evaluate the Shyft modelling framework in its ability to simulate primary water balance components in the Budhi Gandaki catchment, with its complex terrain and limited availability of data. To achieve this goal the study will be divided into the several objectives.

The first will be to investigate spatial and temporal variation in the forcing data (temperature, precipitation, global radiation, relative humidity and wind speed). The recently added bias-adjusted ERA5 reanalysis data is chosen for this study because it has shown to have a lower mean absolute error and higher correlation with observed precipitation than the WDFEI reanalysis. The analysis of temporal variation is done to see if there are any trends or seasonality that might affect the hydrology of the High Mountain Asia and the Budhi Gandaki catchment. This is important because potential changes in the climate may affect the model performance. The spatial variation will be investigated through seasonal maps of the forcing variables and by statistical methods. The analysis of spatial variation may help explain if the forcing data is able to capture leeward and windward precipitation patterns that might affect the model performance. Furthermore, two different interpolation methods in Shyft (Bayesian Temperature Kriging and Inverse Distance Weighting) will be discussed.

The second objective is to evaluate the effect of different terrain representations and snow routines on model calibration and performance using observed river discharge as target. To achieve this, the catchment is discretised using two different discretisation methods. A regular grid representation will be compared with four different Triangular Irregular Networks (TINs) of different resolutions. The two snow routines used include the Gamma Snow routine (simplified surface energy balance model for snow melt) and the recently introduced Snow Tiles routine. The Gamma Snow routine is a simplified surface energy balance model that has been tested in the region before with satisfactory results. The Snow Tiles routine is a temperature-index based model which has never been tested in the region. The Snow Tiles routine is chosen because it has considerably fewer parameters than the Gamma Snow routine, potentially making calibration faster and thus lower uncertainty stemming from equifinality. The model performance will be evaluated by comparing simulated and observed discharge using visualisation and evaluation criteria such as the Nash-Sutcliffe Efficiency (NSE), Kling-Gupta Efficiency (KGE) and Root Mean Square Error (RMSE).

The third objective is to evaluate the impact of the different terrain representations and snow routines on simulated snow-covered area, snow water equivalent, snow cover duration and glacier melt. Snow melt can be a significant contributor to runoff, and it is therefore important that the accuracy of modelled snow is assessed. Simulated snow-covered area (SCA) is compared with observed snow-covered area and snow-cover duration (SCD) derived using observed SCA by MODIS. This method provides an additional way of evaluating model performance. The evaluation criteria used for this comparison is the Critical Success Index (CSI).

References are provided in: Skavang, J. (2023) Assessing the Shyft Modelling Framework in Nepal: Impact of Snow Routines and Terrain Representation on Simulated Water Balance Components. University of Oslo.

1.5 Expected date for project end
Please choose a date at which you expect your Master thesis project will be finished.
15.05.2023

2. Data description
Here we will ask you some questions about the physical and/or digital data you will study. With the term 'Data', we mean any kind of physical or digital form of data that is used and/or produced during your thesis work.

2.1 (Re)-Use existing data
Will you reuse existing data produced/collected by others?
Yes
Describe the data set that you have obtained for reuse. Typical information needed: type of data, who produced or collected this data, when was this data collected, how did you acquire this data. In case of digital data, what is the size of the data set, what format is used for this data set.

The WATCH Forcing Data methodology applied to ERA5 data set (WFDE5), is a meteorological forcing dataset for land surface and hydrological models. The dataset includes a bias-corrected reconstruction of eleven near-surface meteorological variables derived from the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis (Cucchi et al. 2020). The data set is derived from the ERA5 reanalysis product that has been re-gridded to a 0.5° x 0.5° resolution. The data has been adjusted using an elevation correction and monthly-scale bias based on Climatic Research Unit (CRU) data (for temperature, diurnal temperature range, cloud-cover, wet days number and precipitation fields) and Global Precipitation Climatology Centre (GPCC) data (for precipitation fields only). Furthermore, correction has been done for varying aerosol-loading, and separate precipitation gauge observations. The dataset covers the period 1979-2019, and has a hourly temporal resolution. The dataset is distributed through the Copernicus Climate Change Service (C3S) Data Store as monthly files in netCDF format. The netCDF format is self-describing data that machine-independent. The format supports creation, access, and sharing of array-oriented scientific data. The data is downloaded at https://cds.climate.copernicus.eu/cdsapp#!/dataset/10.24381/cds.20d54e34?tab=form. The WFDE5 data is distributed under the CCA 4.0 License.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H., Hersbach, H., and Buontempo, C.: WFDE5: bias-adjusted ERA5 reanalysis data for impact studies, Earth Syst. Sci. Data, 12, 2097–2120, https://doi.org/10.5194/essd-12-2097-2020, 2020.

The topographical and land cover data sets used in this project includes the SRTM 1 Arc-Second Global from the NASA's Shuttle Radar Tomography Mission (NASA-SRTM) that provides a digital elevation model (DEM) with approximately 30 meters resolution. The DEM is on the GeoTIFF/TIFF format and can be downloaded at https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1?qt-science_center_objects=0#qt-science_center_objects.

Earth Resources Observation And Science (EROS) Center (2017) Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global. U.S. Geological Survey. DOI: 10.5066/F7PR7TFT

Furthermore, the Moderate Resolution Imaging Spectroradiometer (MODIS)/Terra Snow Cover Daily L3 Global 500 m SIN Grid v6.1 dataset was used. The MODIS product MOD10A1 provides snow cover, snow albedo and quality assessment (QA) with a spatial resolution of 500 m x 500 m. The data has a daily temporal resolution. The MOD10A1 covers the period 24 February 2000 to present. The Terra MODIS snow products have been validated under both ideal and non-ideal conditions (Aalstad et al. 2020; D. K. Hall and Riggs 2007). The MODIS instrument onboard the Terra (2000-present) and Aqua (2002-present) satellites provides multispectral imagery in the visible shortwave infrared (VSWIR) wavelength spectrum (Aalstad et al. 2020). The optical sensors measures the reflected shortwave radiation from the upper 5-10 cm in the snowpack in multiple

bands of the VSWIR sprectrum. These bands can be combined to give a spectral signature of the snow so that snow-covered area, albedo, snow grain size and and impurity concentration can be measured (Aalstad et al. 2020). MODIS has a daily revisit period with a a ground sampling distance (GSD) of 500 meters (Aalstad et al. 2020). The NASA MODIS snow cover products compare favourably with other products in terms of quality and resolution (D. K. Hall and Riggs 2007). The snow-cover is found using the Normalized Difference Snow Index (NDSI) and a series of screens designed to alleviate errors and flag uncertain snow-cover detections (NSIDC 2023). The datasets come in HDF-EOS formatted data files, and can be downloaded at https://modis.gsfc.nasa.gov/data/dataprod/mod10.php. The MODIS data provided here is pre-processed by a method described by Aalstad, K. et al. (2020), and includes snow-covered area, snow cover duration, mean snow-covered area and mean snow cover duration for the Budhi Gandaki catchment in the period 2000-10-01 to 20015-09-30. The pre-processed data has a daily resolution and sinusoidal projection, and a netCDF format. Only data from the Terra satellite is being used (MOD10A1). The NASA snow-cover product is distributed by the National Snow and Ice Data Center (NSIDC) (D. Hall et al. 2006; D. K. Hall, Riggs G., et al. 2015). The fSCA and fSCD used in this study, is retrieved from MODIS using methods described in (Aalstad et al. 2020). The pre-processed data set contains fractional snow-covered area for each pixel in the domain with a daily resolution, as well as statistics such as mean fSCA and snow-cover duration (fSCD). The data set includes the water years 2000/2001-2014/2015 (2000-10-01 to 2015-09-30). The fSCA value range from 0 to 1, where 1 means that the pixel is completely snow-covered and 0 means not snow-covered. The fSCD is the sum of daily fSCA within a given water year. A fSCD equal to 365 indicates that the pixel is snow-covered for the entire water year (i.e. a glacier), while a fSCD equal to 0 indicates that a pixel is never snow-covered. The projection of the pre-processed and native MODIS data is sinusoidal. The pre-processed MODIS data has not been published before, but is distributed under the CC BY 4.0 License through this work.

Hall, D. K., G. A. Riggs, and V. V. Salomonson. 2006. MODIS/Terra Snow Cover 5-Min L2 Swath 500m. Version 5. Boulder, Colorado USA: NASA National Snow and Ice Data Center Distributed Active Archive Center. http://dx.doi.org/10.5067/ACYTYZB9BEOS.

The Land Cover Classification System (LCCS) Land Cover Map Fine Resolution V2.3 Global dataset from the GlobCover Portal provides global composites and land cover maps. The GlobCover products have been processed by the European Space Agency (ESA) and by the Université Catholique de Louvain using input observations from the 300 m MERIS sensor on board on the ENVISAT satellite mission. The land cover maps covers December 2004 - June 2006 and January - December 2009. The surface reflectance mosaic products are projected in a Platé-Carré projection (WGS84 ellipsoid) with a 1/360° pixel resolution. The land cover classes as defined by a set of classifiers. The data format is ".tif", and is downloaded from: http://due.esrin.esa.int/page_globcover.php. No license is provided, but the data may be used for educational and/or scientific purposes without any fee on the condition that ESA is credited, and the Université Catholique de Louvain is used as the source of the GlobCover product.

ESA & UCLouvain (2010) GlobCover 2009. ESA. http://due.esrin.esa.int/page_globcover.php

A shapefile for the Budhi Gandaki catchment is used to make maps. This shapefile contains the catchment outline, catchment id and the catchment area. The file is the same the data set used in Bhattarai 2020. There were no license information available for the shape file. The file was allowed to share.

River network data is used for the catchment map. The river network data is a vectorised line network of all the global rivers that have a catchment area of at least 10 km of a river flow of more than 0.1 m3/s (Lehner and Grill 2013). HydroRIVERS is a free and open-source database for scientific, educational and commercial use. The downloaded zip data contains a documentation file (.pdf) and a shapefile that can be visualised in QGIS. Version 1.0 of HydroRIVERS is downloaded. The HydroRIVERS is freely available at www.hydrosheds.org for scientific, commercial and educational use. The data is distributed under a license described in https://

data.hydrosheds.org/file/technical-documentation/HydroSHEDS_TechDoc_v1_4.pdf in Appendix A: “Hydrosheds version 1 - License Agreement”.

Lehner, B., Grill G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world’s large river systems. Hydrological Processes, 27(15): 2171–2186. **https://doi.org/10.1002/hyp.9740**

The regular grid cell data produced by Bhattarai 2020, and is shared under the CCA4.0 License at: https://zenodo.org/record/3567830#.ZHXsZS9Bzuw.

TINs for the Budhi Gandaki catchment are made using the Rasputin software. The TINs are created using the Rasputin software (Silantyeva et al. 2023). The Rasputin software can convert a point set of coordinates to a TIN using a Digital Elevation Model (DEM) and an outline of the catchment in a wkt-file (Bhattarai, Silantyeva, et al. 2020) using Delaunay triangulation methods (Silantyeva et al. 2023). This is done by converting a DEM into simplified triangulated meshes. The triangulation routine in Rasputin is based on the CGAL Delaunay method (Silantyeva et al. 2023; Yvinec 2023). The 2D Delaunay triangulation in Rasputin is non constrained, meaning that the Delaunay triangulation is purely based on the position of a given set of vertices disregarding how they ought to be connected by edges (Silantyeva et al. 2023). The land types are designated to each TIN by determining the position of the middle point in the land cover data set. In this study, Rasputin uses the GlobCov 2009 data set with a 300 m resolution for the land types. Only one land type is assigned to each triangle, which differs from the regular grid cells that have fractional land cover types (Silantyeva et al. 2023). In addition to creating the TINs, the software also calculates physical parameters for each TIN facet (such as slope, aspect and area) from the geometry (Bhattarai, Silantyeva, et al. 2020). The TIN generated from Rasputin is in a XDMF format and can be visualised using various visualisation tools or in a H5 format that can be parsed and visualised using Python. In this study, the H5 format will be used to parse the TINs for Shyft and to visualise the TINs. The Rasputin software is freely available under GNU GPL v3.0 license. The TINs are not published before, but not shared as a part of this work under the CC4 License at https://github.com/jacobqs/MST-Himalaya/tree/main/shyft_workspace/shyft-data/budhi_gandaki/tin_archive as a “.h5” file and “.xdmf” file.

GeoJSON polygons of countries are downloaded from datahub.io. These polygons are used in figures where country boundaries are illustrated. The data is in the GeoJSON format and is licensed under the Open Data Commons Public Domain Dedication and License (PDDL) v1.0.

Glacier polygons are downloaded from the NASA Earth Data Explorer in a shapefile format (Li et al. 2021). The glacier polygons are globally glacier polygons based on the Ralph Glacier Inventory (RGI) v6.0. The RGI glacier polygon data is provided under an CC BY 4.0 License at https://www.glims.org/RGI/

RGI Consortium, 2017. Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 6. [Indicate subset used]. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. doi: **https://doi.org/10.7265/4m1f-gd79**

Daily stream flow data from the period 2000-2015 is obtained from The Department of Hydrology and Meteorology, DHM (DHM 2022). The flow gauging station is located at Arguhat bazaar (485 m.a.s.l.) at 28.043611 N and 84.816389E. It has a drainage area of 4270 km2 (DHM 2022). The station has been operating since 28 November 1963. To get observed river discharge data, please contact Innovation PostDoctoral Researcher Olga Silantyeva (University of Oslo) at olga.silantyeva@geo.uio.no.

For more info, please refer to: Skavang, J. (2023) Assessing the Shyft Modelling Framework in Nepal: Impact of Snow Routines and Terrain Representation on Simulated Water Balance Components. University of Oslo.

2.2 Data collection
Will you collect raw data yourself?
No

2.3 Data analysis
After data collection and/or getting acquainted with your acquired existing data set, it is time to analyse your data and these analyses will again produce new data. How will you analyse your data? Describe the workflow. Do you perform lab experiments? Do you use any specific software to analyse your data? What kind of data is produced from these analyses? Most of the data produced by analyses is probably stored as digital data. What format do you use and what is the anticipated size of the analysed data set? Consider carefully the structure of your data set and and how you name your files and directories.

In the project I have used Pyhton for pre-processing and analysis of data. I have also used QGIS for processing spatial data. The Shyft platform is a Python package that provides spatially distributed conceptual hydrological models. It is an optimised platform for the implementation of well known hydrological models. The main model forcing variables include atmospheric temperature [°C], precipitation [mm/h], relative humidity [-], wind speed [m/s], incoming shortwave radiation [watt/m2]. The main variables for evaluation include discharge [m3/s] and snow-covered area [-]. The netCDF format provides both the data and the metadata related to these variables.

The first step will be to prepare the data such that it can be used in Shyft. The pre-processing of data includes handling missing data (if any), outliers and unexpected data values.
For the exploratory analysis, I have visualised the data using Pandas and Matplotlib. To get a description of the data I have used a statistical analysis including the Pandas, Seaborn, Xarray and Numpy packages. These packages will help generate descriptive statistics that summarises the central tendencies, dispersion and shape of a data set's distribution.

Past climate is explored using the WFDE5 reanalysis of a 30 year period (1990-2019). In this analysis I have explored if there are any temporal and spatial patterns in the near-surface variables.

Digital Elevation Model (DEM) data will be used to make maps and to obtain the elevation for the WFDE5 data. The catchment delineation process consists of several steps. First the DEM files need to be merged.

Shyft is an open-source cross-platform hydrologic modelling toolbox. The Shyft API provides all functionalities needed to construct a model, calibrate it, and run it. The Shyft software is distributed under the GNU LGPL version 3.0 at https://gitlab.com/shyft-os/shyft. To compile and run Shyft, several requirements need to be met. These are described at https://gitlab.com/shyft-os/shyft or in Skavang, J. (2023) Assessing the Shyft Modelling Framework in Nepal: Impact of Snow Routines and Terrain Representation on Simulated Water Balance Components. University of Oslo.

Burkhart, J. F. et al. (2020) Shyft v4.8: A Framework for Uncertainty
Assessment and Distributed Hydrologic Modelling for Operational Hydrology, Geosci. Model Dev., https://doi.org/10.5194/gmd-2020-47

The workflow using Shyft consists of several steps. The first step configure each model. This includes configuring the simulation, model, region domain, interpolation methods and calibration. The model parameters will define how the model equations will respond to a certain meteorological forcing. These parameters are specific to the model type, and can be obtained directly from the model class. The parameters can be divided into region parameters (all cells in the same domain have the same parameters) and catchment parameters (specific to the (sub)catchment). The fourth step is to construct the region model. The region model combines information about the model domain, or region, which is provided by the cell, and the model we want to use. The link between a region and the model are the region (or catchment) specific model parameters.

In step five the forcing data is fed to the region model by constructing a region environment. This is done by constructing forcing data time series that are geolocated (so called “sources”). To create a time series of forcing data, a time axis is defined, which is then combined with the data values. In order to add sources to the region model, collections of sources of a certain type (such

as temperature and precipitation) are organised as SourceVectors. These SourceVectors are then added to the regional model.

The sixth step is to run the model. To do this initial values for the state variables need to be defined. Furthermore, since the model-stack is executed cell-by-cell, the source variables (e.g. model forcing) need to be interpolated from the source locations to the cell locations, such that each cell holds a complete set of forcing variables. These are used to run the model cell-by-cell.

The nest step, is to calibrate the model. Each configuration is calibrated daily for 2000-01-01 to 2004-12-31 using the Min BOBYQA optimisation method.

Lastly, the simulation is ran again with optimised parameters and the simulation results are extracted. The model results are stored in memory (and not written to a file) and can be accessed via the region model. The results can either to obtained cell-by-cell or aggregated at the catchment level.

3. Documentation, metadata and organization

Consider your final data set to be reused by someone else. What does he/she need to know about your data set so he/she can start using it right away? How would you like the data to be organised?

For more information, a good starting point is the online MANTRA - Research Data Management Training course. Take your time to walk through these free online modules on good practice in research data management.

3.1 Documentation

Provide information on how you will add documentation to your data set and what information you will provide.

Think about content of your files, file names, how to use the data set (workflow), total data set size, formats used, etc.

MANTRA - Documentation, metadata and citation

This project will use a scripted language (i.e. Python) that can be documented to perform analysis and save the outputs in a separate file/variable/data frame. Jupyter Notebook is in particular well suited for documenting the workflow of the data analysis process, while regular Python scripts may be better suited for building packages containing useful function and classes that can be imported into the Jupyter Notebook environment.

A Github/Gitlab repository will be used as the end-user documentation for source code of the project. This repository will include a “README.md” file that contains the source code, and information about what the study is set out to do, how it contributes new knowledge to the field, what the research questions are, what methodologies are used, how to get help with the repository, acknowledgements, and information about necessary software/packages needed to run the code. Furthermore, it will contain information about how the data can be accessed and how the data is processed. Additionally, it will contain information about the licence for using the data.

Documentation for the different datasets will provided in a Technical Documentation file in the data folder at Zenodo. This documentation will contain the location of the data, and information about the data variables and units, methods used to generate the data, information about the format, parameters, and temporal and spatial coverage. It will also provide links to the data providers.

Documentation about the processing of geospatial data in QGIS is given in the Master Thesis appendix (please see: Skavang, J. (2023) Assessing the Shyft Modelling Framework in Nepal: Impact of Snow Routines and Terrain Representation on Simulated Water Balance Components. University of Oslo).

3.2 Metadata

Metadata will increase the usability of your data set considerably. Metadata is "data that provides data about other data". In other words, it is "data about data".

To learn more about metadata and metadata standards:

https://en.wikipedia.org/wiki/Metadata
For geospatial metadata:
https://en.wikipedia.org/wiki/Geospatial_metadata
If your data is stored with metadata, what metadata standard is used? Where is the metadata stored, in the actual data file, or as separate metadata files?
MANTRA - Documentation, metadata and citation

Descriptive metadata are common fields such as title, author, subject, contact information, funder and keywords that help users discover online sources through searches and browsing. The descriptive metadata will be stored as a research object at RoHub (https://reliance.rohub.org/). The research object also contains information about the study area, related literature, contributors, publisher, copyright holder and licence. Additionally, a general-purpose open repository, Zenodo, is used to store data. Zenodo will also mint a persistent digital object identifier (DOI) for each submission, which makes items easily identifiable.

Metadata is also stored within netCDF files and shapefiles. Data in the NetCDF format is self-describing, portable, scalable, appendable, sharable and archivable. The global metadata stored within the netCDF files contain information about when the file was created, name of the institution and/or model used to generate the file, links to peer-reviewed papers and technical documentation describing the climate model and software used to generate the file. Moreover, the netCDF format contain metadata about variable dimensions (time, latitude, longitude and height) and variable metadata (e.g. units, averaging period (if relevant) and additional descriptive data).

The shapefile format is a geospatial vector data format for geographic information system (GIS) software. The coordinate data assumes a Cartesian coordinate system, using (X, Y) or (Easting Northing). Each file usually has an attribute that describe it such as name.

3.3 Organisation of the data

How is you data organised? Consider carefully the structure of your data set. This will also considerably increase the quality of your data set.
MANTRA - Organising data

The project will be organised in a Github/Gitlab repository. The repository will be organised with a “README.md” file that contains all the necessary end-user documentation, a “.gitignore” file that contains files that are ignored by Git, a “shyft_workspace” folder containing the code for Shyft simulations, and a “LICENCE.md” file that explains the legal licensing.

All of the data will be stored at Zenodo. The data will be organised in a “Shyft” folder containing raw data and processed data used for Shyft simulations. All other data that is used for other purposes, mainly analysis, is stored in the “Analysis” folder.

4. Storage and backup during the research process

MANTRA - Storage and security

4.1 Data storage

Provide some information where your data is stored during your master thesis project. If your project requires specific storage resources, please contact drift@geo.uio.no. Please provide information about what kind of resources you need and the name of your supervisor.

The data will be stored at Zenodo with a citable digital object identifier (DOI) at https://zenodo.org/record/7992195. The Zenodo DOI is 10.5281/zenodo.7992195.

4.2 Data backup

To be sure your data is not lost, it is recommended to make a back up of your data regularly. If you store data on your home-directory or on the data server of the department, your data is securely taking backup of on a regular basis.
Read more about backup at UiO here:
https://www.uio.no/english/services/it/store-collaborate/backup/

If your data is not stored on the above mentioned areas, please provide information on how you plan to take backup of your data.

Zenodo has a 12-hourly backup cycle for metadata with one backup sent to tape storage once a week.

5. Legal and/or ethical issues
Are there any legal and/or ethical issues related to your data? Does your data need specific protection, because of e.g. commercial value or GDPR issues? If you are (re)-using data provided my third parties, do they have any requirements regarding the use of these data?
The data I will use is free and publicly available, and can only be used for educational purposes.

The requirements regarding the use of the data include that:
- I will hold no individual(s), organization(s), or group(s) responsible for any errors in the models or in their output data
- I acknowledge the potential limitations of the data obtained, and that although the forcing data has been subjected to a quality control procedure, unrecognized errors almost certainly remain
- the downloaded data is publicly available, but that I cannot sell the data or use the data for commercial purposes
- any restrictions of the data are detailed in an assigned license
- I restrict my use of the data to the assigned license
- I will appropriately credit the data providers, either by citing the DOIs relevant to the project or by an acknowledgement
- I will provide the associated reference of my Master Thesis to the science team behind the data
- I will not copy or download items without first obtaining a license from the rights controller. If copyrighted, permission should be obtained from the copyright owner before use. If not copyrighted, material may be reproduced and distributed without further permission, but make sure that this material is properly cited.
- results from non-commercial research are expected to be made generally available through open publication and must not be considered proprietary
- I will respond to requests from data providers for feedback about my results in order to aid participation groups in understanding and improving their models
- There are no ethical issues
- No sensitive data in the thesis
- All the related licenses are acknowledged

Furthermore, the data from the Department of Hydrology and Meteorology, Government of Nepal is restricted to a person/organization, and can only be used for sole purpose of this project's work/thesis. To obtain the observed discharge data, please contact Olga Silantyeva at olga.silantyeva@geo.uio.no.

6. Data sharing and long term preservation
Here we ask you to consider what you want to do with your data after you have finished your master thesis project. Before answering this question, discuss this with your supervisor(s). For more information, you can read 'Master Student Clearance - What to do with your data?'
6.1 Data sharing
Will any of the physical and/or digital data supporting the thesis (e.g. organised project archive folders with images, drawings, spreadsheets, databases, samples etc) be made available to others on request or open access (e.g. to the host project, research lab/community, museums, open-access web-based organization or national research data archive)?

Yes

6.2 Data repository
If you study physical data (rock samples, ice cores, water samples, a.o.), where will these samples be stored/archived after you have finished your project? Are they part of a larger collection connected to an institution?

If you plan to archive digital data, describe here which platform you intend to use for archiving your data. Some Open Access possibilities are:
NIRD's Research data Archive; The Norwegian e-Infrastructure for Research Data (recommended)
DataverseNO; an archive service for open research data, drifted by UiT The Arctic University of Norway. For UiO users, data sets up to 1GB
Open Science Framework; Center for Open Science. A collaboration platform to increase openness, integrity and reproducibility of research
Zenodo; Developed by the European OpenAIRE program and run by CERN. Data sets up to 50GB.
GitHub; Suitable for sharing open source code. Not suitable for sharing large data sets
Other possibilities could also include databases of partner institutions, but it should be possible for the university to extract the data when needed without any costs.
Some ways NOT to archive your data are:
A CD-ROM/External hard-drive/USB-Stick. This type of storage media have tendencies to become obsolete over time, get lost or get broken and is probably not very accessible for others.
Undocumented on the network drive of the department. This data will be deleted after you leave the university.
What data formats do you intend to use for archiving your data? For data to be re-usable, it is important to either use standardized data formats (like netCDF, SEGY, Shapefiles, etc) or ASCII formats (that are human readable) when archiving your data. Is it necessary to convert your produced data to more re-usable data formats?

The research data will be shared according to FAIR principles (Findable, Accessible, Interoperable and Reusable). The FAIR principles are laid out in The FAIR Guiding Principles for scientific data management and stewardship' (Wilkinson et al, 2016).

Findable means that the metadata and the data should be easy to find for both humans and machines. This includes:
- F1. Data and metadata are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. Data and metadata are registered or indexed in a searchable resource

Accessible means that the user finds the required data, and that she/he knows how the data can be accessed. This includes:
- A1. Data and metadata are retrievable by their identifier using a standardised communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Data and metadata are accessible, even when the data are no longer available

Data usually needs to be integrated with other data. Interoperable means that the data need to interoperate with applications or workflows for analysis, storage and processing. This includes:
- I1. Data and metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation (such as standard metadata vocabularies or ontologies)
- I2. Data and metadata use vocabularies that follow FAIR principles (such as generic or discipline-specific metadata standards)
- I3. Data and metadata include qualified references to other data and metadata. (Related objects are linked using persistent identifiers and citation.)

The ultimate goal of FAIR is to optimise the reuse of data. Therefore, metadata and data should be well described so that they can be replicated and/or combined in different settings. This includes:
- R1. Data and metadata are richly described with a plurality of accurate and relevant attributes
- R1.1. Data and metadata are released with a clear and accessible data usage license
- R1.2. Data and metadata are associated with detailed provenance
- R1.3. Data and metadata meet domain-relevant community standards

The project has a Research Object (RO) on RoHub (https://reliance.rohub.org/) which will function as a framework for packaging all the research products into FAIR objects. An RO is a multidimensional digital object that includes essential information about experiments and

investigations to facilitate reusability, reproducibility and better understanding. The RO can aggregate an arbitrary number of heterogeneous resources, internal or external (linked by reference), such as the data used or results produced, methods that are used, and people that are involved. The research object can include any number of metadata associated to the resources (or the RO itself), enabling the understanding and interpretation of the scientific work. The RO can be found at https://reliance.rohub.org/648630a1-c6a3-4f64-8a39-e4cd29a55221?activetab=overview

GitHub/Gitlab will be used for code sharing, version control and collaboration. Git is a free and open source distributed version control system designed to handle everything from small to large projects (https://git-scm.com/). The Git software can track any set of files, such that there is a record of everything that has been done. Git also allows the user to revert to specific versions whenever there is a need for it. Collaboration is east with Git allowing changes from multiple users to be merged into one source. Git also allows branching which lets the user branch out of the original code base and work with features on a separate branch that can be merged back to the main branch. This lets the user more easily work with other people and gives a lot of flexibility in the workflow. Files that are tracked by adding them to the staging area. To record the changes to a files, the staged files must be committed. A commit contains metadata about when the commit is done, by who and a description about what change is done. The purpose of the staging area is to keep track of the files, but only the files that are ready for a commit will be committed. Github/ Gitlab stores a copy of the Git repository online. This lets the project to be stored at a centrally located place where users can upload (push) changes and download (pull) changes. This lets the users collaborate more easily. All the commits can also be seen here.

The Zenodo repository is linked to GitHub/Gitlab to generate a citable Digital Object Identifier (DOI) for the repository.

7. Links to thesis and data set
Here we ask you to fill in links to your thesis and/or data set. These can only be filled in after you have submitted these. Link to your data set is only visible if you have answered 'Yes' on question 6.1.
Link to your data set
If you have archived your data set in any kind of repository, please enter a link to your data set here (e.g. a DOI).

https://reliance.rohub.org/648630a1-c6a3-4f64-8a39-e4cd29a55221?activetab=overview

DUO-link to your Master thesis: