Description and setup of the ML models for the estimation of Snow Water Equivalent (SWE) and runoff in Alpine watersheds

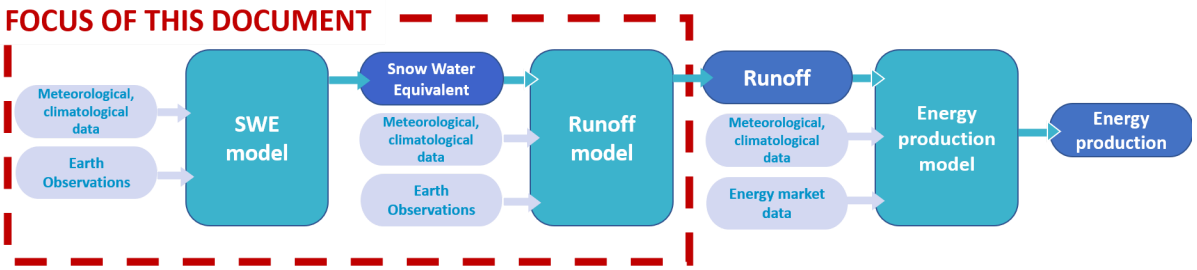
The present document describes the datasets and setup of the ML models used for the estimation of Snow Water Equivalent (SWE) and runoff in Alpine watersheds within the scope of SnowPower, one of the projects selected in the **I-NERGY 1st Open Call** (https://i-nergy.eu/).

This document is structured as follows:

1. The SWE and runoff models in the context of SnowPower
2. SWE model
3. Runoff model

1. The SWE and runoff models in the context of SnowPower

SnowPower is a Software as a Service (SaaS) for the monitoring and forecasting of hydropower generation in the Alps, on a domain including watersheds located in France, Italy, Switzerland, and Austria. SnowPower focuses on the estimation of the amount of water expected to flow into the watersheds of interest (runoff) through the assessment of the Snow Water Equivalent (SWE) within their boundaries.
The software seamlessly integrates satellite, in-situ data, reanalysis products and seasonal forecasts with three Machine Learning (ML) modules–Snow, Runoff, and Electricity respectively–that operate sequentially with the goal of providing reliable and accurate data-driven predictions. Output of the SnowPower solution is, in fact, not only hydropower energy generation, but also, and perhaps most importantly, the variables driving it (SWE and runoff), as these represent crucial pieces of information for energy planning and management decisions at power plant level.



A schematic diagram of the modules of SnowPower.

The three modules of SnowPower and their interconnections are briefly introduced below.

Snow: SWE is an important descriptor of the amount of water stored in the snowpack, as it represents the water resource that will be potentially exploitable for energy generation during the melting season. Nonetheless, SWE on-site measurements are often not automated,

resulting in data that are sparse in both space and time. Furthermore, large-scale available observations can be incomplete (e.g. the Copernicus SWE product is masked in mountain regions that are focal to SnowPower). Through the Snow module, SnowPower overcomes these limitations by relying on a targeted ML model for SWE estimation (described in 2). The outputs of the Snow module are then fed to the Runoff module.

Runoff: water runoff in a watershed, both due to precipitation or snowpack melting, is a significant proxy of water availability for hydropower generation. In the Runoff module, SWE combined with meteorological and orography data serve as input to a tailored ML model to predict runoff in Alpine watersheds, as detailed in Section 3. The outputs of the model are then used as inputs to the Electricity module.

Electricity: in this module, the estimation of the hydropower generation is informed by the estimated runoff and a set of variables describing energy demand, such as temperature and market demand. Here, a ML approach developed by De Felice et al.[1] for photovoltaic production is used as a reference for estimating electricity production, in which open data about hydropower energy production play an important role in improving the quality of the model's results.

This document is aimed at describing the Snow and Runoff modules only.

2. SWE model

The SWE model is at the core of the Snow module of SnowPower. The SWE model yields information about SWE at watershed level by feeding it data about each watershed's orography and meteorology, as detailed below.

2b. Feature selection and feature engineering

Several variables have been tested and selected as meaningful features for the estimation of the SWE, and can be found in the table below grouped by data type.

Data type	Variable name	Description	Source
Satellite-derived	Snow Cover Extent	Percentage of area (commonly a cell of the data grid) that is covered by snow. The product selected is based on MODIS satellite data and provides daily gridded data over Continental Europe starting from March 2017.	500m Snow Cover Extent (SCE) version 1 by the Copernicus Global Land service[2]
Measurement-derived	North-Atlantic Oscillation (NAO)	A north-south dipole of anomalies. It is an indicator of large-scale meteorological conditions: NAO is associated with changes	Daily NAO index from NOAA[3]

[1] De Felice, M., Petitta, M., & Ruti, P. M. (2015). Short-term predictability of photovoltaic production over Italy. Renewable Energy, 80, 197–204. https://doi.org/10.1016/j.renene.2015.02.010

[2] Copernicus Global Land service, (2017) Snow Cover Extent (SCE500)-Continental Europe (CEURO)-500m. The product was generated by the land service of Copernicus, the Earth Observation program of the European Commission. The research leading to the current version of the product has received funding from various European Commission Research and Technical Development programs. The product is based on SCE500-CEURO-500m data ((c) ESA and distributed by ENVEO).

		in zonal and meridional heat and moisture transport and with the intensity and location of the North Atlantic jet stream.	
Reanalysis	Snow depth	Amount of snow within a grid cell, in meters of water equivalent, i.e. the depth the water would have if the snow melted and was spread evenly over the whole grid box.	ERA5[4]
Reanalysis	Precipitation	Accumulated liquid and frozen water, comprising rain and snow, falling to the Earth's surface. Provided as hourly values.	ERA5
Reanalysis	10m v- and u-component of wind	Horizontal speed of air moving towards the east and towards the north, respectively, at a height of ten meters above the surface of the Earth, in m/s.	ERA5
Reanalysis	Temperature	Temperature of air at 2m above the surface of land, sea or inland waters	ERA5
Reanalysis	Surface net solar radiation	This parameter is the amount of solar radiation (both direct and diffuse) that reaches a horizontal plane at the surface of the Earth minus the amount reflected by the Earth's surface in J/m^2.	ERA5
Orography	Digital elevation model	Represents the surface of the Earth in terms of height of land features. Here we use the GLO-30 instance of the Copernicus DEM.	ESA[5]
Geometry	Basin area and shape	Georeferenced polygons from the HYDROSHEDS-HYDROBASINS product.	HYDROSHEDS[6]

Not only some of these features require pre-processing (e.g. resampling values along the time dimension to obtain daily values), but some can be further engineered to provide brand new, meaningful features. In particular, the precipitation is processed to obtain the 3-month Standardized Precipitation Index (SPI3, Mckee et al.[7]); the daily minimum, maximum, and mean temperature are extracted from the ERA5 hourly values; the mean daily wind speed is calculated from the two horizontal components, and minimum, mean, and maximum altitude of each watershed are extracted from a digital elevation model.

It is important to remark that all the aforementioned features are georeferenced (except for NAO, that is an index), and thus their spatial distribution must be properly accounted for when preparing the features to feed to the model. For this reason, the selected variables need proper aggregation over the Alpine watersheds of interest when finalizing the feature dataset.

---

[3] https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml

[4] Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2018): ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed in July 2022), 10.24381/cds.adbb2d47

[5] https://doi.org/10.5270/ESA-c5d3d65

[6] Lehner, B., Verdin, K., Jarvis, A. (2008): New global hydrography derived from spaceborne elevation data. Eos, Transactions, 89(10): 93-94. Data available at https://www.hydrosheds.org

[7] Mckee, T. B., Doesken, N. J. & Kleist, J. The relationship of drought frequency and duration to time scales. 17–22 (1993)

The final set of features for the SWE model is thus composed of: Snow cover extent, snow depth, daily precipitation, SPI3, minimum, mean, and maximum temperature, wind speed, surface net solar radiation, minimum, mean, and maximum altitude of the watershed, NAO, and day of the year.

2c. Model setup and training

Model training is constrained by data availability: we train our model using as target the daily SWE data for the period October 2021 - May 2022 for some major watersheds in the Western Alps, provided by MobyGIS s.r.l.[8]. Therefore, we train our model on a subset of the area covered by the SnowPower solution: this aspect is further detailed in 2d below.
The features mentioned in 2b are thus coherently preprocessed to match the time period and area covered by the available target data: furthermore, they are standardized before being fed to the model. A train-test split procedure is used to provide a testing dataset.

The selected algorithm is the boosted trees regressor encoded in the XGBRegressor class of the XGBoost[9] Python library. We perform hyperparameter tuning using K-fold Cross-Validation to identify, among others, the optimal number of trees, their maximum depth, subsample size per tree, and learning rate.

2d. Model performance

The model reports on the testing dataset a mean absolute percentage error (MAPE) of about 14%.
Note that the data we can use as target covered the Western Alps in October 2021-May 2022 only: we thus validate our model by applying it on each of the Alpine watersheds included in the SnowPower solution for the snow accumulation seasons (i.e. October to March) 2017-2020, and by comparing model predictions with the snow depth from ERA5, which was also included among the features. We are in fact limited by the scarcity of SWE data over the Alps. Considering the difference between the two variables, we choose not to calculate conventional performance indicators (e.g. RMSE, or percentage error) to compare SWE resulting from our model with snow depth, but look instead at the temporal patterns and correlation between the two time series for each watershed. This comparison yields satisfactory results, with a R^2 >0.9 with p-values well below 0.05 for more than 65% of the basins.
The resulting daily SWE data for 2017-2020 are then used as inputs for the Runoff model.

3. Runoff model

The runoff model within the homonymous module is tailored to provide the runoff in each watershed during the snowmelt season, which, for the purposes of SnowPower, we define to last from April to September. From this point of view follow some of the choices we take both in feature engineering and in model training/testing, and that are detailed in Sections 3a and 3b respectively.

---

[8] https://www.waterjade.com/it/mobygis/

[9] Chen, Tianqui, and Guestrin, Carlos, (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

3a. Feature selection and feature engineering

Other than the SWE yielded by the Snow module and some key meteorological and orography-related features that are also used in the SWE model (i.e. surface net solar radiation, minimum, mean, and maximum temperature, wind speed, surface net solar radiation, minimum, mean, and maximum altitude of the watershed, NAO, and day of the year), another feature is introduced to describe the abundance/lack of precipitation during the winter snow accumulation season. This is the total cumulative precipitation, computed from ERA5 daily values.

3b. Model setup and training

The training feature dataset encompasses the snowmelt seasons of 2017-2019, while the target is derived from the daily runoff of ERA5 aggregated at basin level.
Again, after testing several algorithms we select the random forest regressor encoded in the XGBRegressor class of the XGBoost[10] Python library, and perform hyperparameter tuning using K-fold Cross-Validation as mentioned in 2b.

3c. Model performance

Model testing performed on the testing dataset reports a MAPE of about 27%, with 8 basins reporting a MAPE lower than 20%.
The MAPE on the validation dataset (April-September 2020) amounts to 44%, and can be quite different from one watershed to another (ranging within 29% to 62%). Nonetheless, the model proves to be quite effective in describing the average behaviors and the peaks: when correlating our results for each watershed with its validation target from ERA5, the 25° percentile of the R^2 value over our domain is about 0.54, and the median is 0.68.

[10] Chen, Tianqui, and Guestrin, Carlos, (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785