README.md

# Learning From Hydrological Models' Challenges: A Case Study from the Nelson Basin Model Intercomparison Project

Mohamed Ismail Ahmed[1,2*], Tricia Stadnik[3], Alain Pietroniro[3], Hervé Awoye[4], Ajay Bajracharya[1,4], Juliane Mai[5], Bryan A. Tolson[5], Hongren Shen[6], James R. Craig[6], Mark Gervais[6], Kevin Sagan[6], Shane Wruth[6], Kristina Koenig[6], Ralamboi Lilhare[7], Stephen J. Déry[7], Scott Pokorny[8], Hank Venema[9], Ameer Muhammad[10], Markkamoh Taheri[3]

[1] Department of Geography, University of Calgary, Calgary, AB, T2N 1N4, Canada

[2] Irrigation and Hydraulic department, Faculty of Engineering, Cairo University, Cairo University Rd, Giza, Egypt

[3] Schulich School of Engineering, University of Calgary, Calgary, AB, T2N 4V8, Canada

[4] Department of Civil Engineering, University of Manitoba, Winnipeg, MB, R3T 5V6, Canada

[5] Department of Civil and Environmental Engineering, University of Waterloo, ON, N2L 3G1, Canada

[6] Water Resources Department, Manitoba Hydro, Winnipeg, MB, R3C 0G8, Canada

[7] Department of Geography, Earth and Environmental Sciences, University of Northern British Columbia, Prince George, BC, V2N 4Z9, Canada

[8] Strategic Systems Engineering, Winnipeg, MB, R3B 1G1, Canada

[9] Associated Engineering (Sask.) Ltd., Winnipeg, MB R3L 2T4, Canada

[*] Corresponding author: Mohamed Ismail Ahmed (mohamedismail.ahmed@ucalgary.ca)

## Data Repository

*This README file gives a brief description of the used datasets (inputs and outputs) of this study. For more detailed explanation, please refer to the journal article.*

## Abstract:

Intercomparison studies play an important, but limited role in understanding the usefulness and limitations of currently available hydrological models. Comparison studies are often limited to well-behaved hydrological regimes, where rainfall–runoff processes dominate the hydrological response. These efforts have not covered western Canada due to the difficulty in simulating that region's complex cold region hydrology with varying spatiotemporal contributing areas. This intercomparison study is the first of a series of studies under the intercomparison project of the international and interprovincial transboundary Nelson-Churchill River Basin (NCRB) in North America (Nelson-MIP), which encompasses different ecozones with major areas of the non-contributing Prairie potholes, forests, glaciers, mountains, and permafrost. The performance of eight hydrological and land surface models is compared at different unregulated watersheds within the NCRB. This is done to assess the models' streamflow performance and overall fidelity without and with calibration, to capture the underlying physics of the region and to better understand why models struggle to accurately simulate its hydrology. Results show that some of the participating models have difficulties in simulating streamflow and/or internal hydrological variables (e.g., evapotranspiration) over Prairie watersheds but most models performed well elsewhere. This stems from model structural deficiencies, despite the various models being well calibrated to observed streamflow. Some model structural changes are identified for the participating models for future improvement. The outcomes of this study offer guidance for practitioners for the accurate prediction of NCRB streamflow, and for increasing confidence in future projections of water resource supply and management.

**Licenses/restrictions placed on the data**: This dataset is made available under a custom license. Please refer to the LICENSE.txt file for the full terms of use.

**Citation (Journal Publication)**: To be updated

**Citation (Data Publication)**: To be updated

This repository contains the following directories:

- 01_study_area
- 02_input_datasets
- 03_verification_datasets
- 04_model_results

A description of the contents of each directory is provided below.

## 01 study area

This directory contains information related to the Nelson-Churchill River Basin (NCRB) study area

### gauges_obs_flows

This directory contains csv files that hold information about the selected observed flow gauges and their metadata. The data for the streamflow gauging stations were derived from the Water Survey of Canada (WSC; https://wateroffice.ec.gc.ca/search/historical_e.html) and United States Geological Survey (USGS; https://waterdata.usgs.gov/nwis/rt) databases for this study. The streamflow data were extracted from the WSC and USGS databases using the *tidyhydat* (Albers, 2017) and *dataRetrieval* (Cicco et al., 2022) R packages, respectively, and were converted to $m^3$/sec.

- *Natural_gauge_stn_Obs_flows_0-1.rar.csv*: observed streamflow values for the 132 selected gauges for the full simulation period (from Jan-2000 to Dec-2017) in $m^3$/sec. Missing values are replaced with -9999.
- *Simulated_Phase_0-1_gauges_by_models.csv*: metadata of the 132 selected gauges for this study along with a flag to show which models simulated the specific gauge (Y). This is Table S10 in the supplemental file.

### shapefiles

This directory contains different ESRI shapefiles of the study area. These shapefiles were obtained by delineating the MERIT Hydro DEM (explained below)

- *NCRB_boundary.zip*: Boundary of the NCRB.
- *NCRB_major_basins.zip*: Major basins of the NCRB.
- *Phase_0-1_subbasins_rivers_gauges.zip*: The selected watersheds' (subbasins') boundary, rivers, and outflow gauges.

## 02 input datasets

A standardized set of geophysical data and meteorological/climate forcing data (provided in this directory) were used by the participating models in the Nelson-MIP study for model setup. The same datasets must be used by future users/contributors to ensure consistency and fair comparison with the published results from the participating models. The datasets were cropped to the NCRB and were aggregated or re-produced in different formats for the convenience of the participating modeling teams.

### GSDE_soil_data

This directory contains the soil texture classification based on the USDA method averaged for the entire soil column (0–2.296 m, used by most participating models) using the GSDE soil data (Shangguan et al., 2014). The GSDE dataset is provided at 30 arc-seconds resolution (~1 km) and was downloaded from http://globalchange.bnu.edu.cn/research/soilw#download. The following table describes the display of the soil texture classification classes in the USDA_soiltexture_combined.lyer (GSDE_NCRB.tif file (legend):

| Cell value (raster .tif) | acronym | description |
|---|---|---|
| 1 | C | clay |
| 2 | SIC | silty clay |
| 3 | SC | sandy clay |
| 4 | CL | clay loam |
| 5 | SICL | silty clay loam |
| 6 | SCL | sandy clay loam |
| 7 | L | loam |
| 8 | SIL | silty loam |
| 9 | SL | sandy loam |
| 10 | SI | silt |
| 11 | LS | loamy sand |
| 12 | S | sand |
| -100 | | NoData or missing data |

### MERIT_hydro_DEM

This directory contains the elevation data of the MERIT Hydro DEM at a ~90 m resolution (Yamazaki et al., 2019) for the NCRB in GeoTiff format. The MERIT Hydro DEM was downloaded from http://hydro.iis.u-tokyo.ac.jp/~yamadai/MERIT_DEM/.

### NALCMS_landcover

This directory contains the landcover data of the NALCMS dataset (Latifovic et al., 2012) for the NCRB in GeoTiff format. The NALCMS dataset was downloaded from http://www.cec.org/tools-and-resources/map-files/land-cover-2010-landsat-30m. The following table describes the display of the landcover classes in the NALCMS_2015_NCRB.tif file (legend):

| Cell value (raster .tif) | description |
|---|---|
| Value 0 | NoData |
| Value 1 | Temperate or sub-polar needleleaf forest |
| Value 2 | Sub-polar taiga needleleaf forest |
| Value 3 | Tropical or sub-tropical broadleaf evergreen forest |
| Value 4 | Tropical or sub-tropical broadleaf deciduous forest |
| Value 5 | Temperate or sub-polar broadleaf deciduous forest |
| Value 6 | Mixed forest |
| Value 7 | Tropical or sub-tropical shrubland |
| Value 8 | Temperate or sub-polar shrubland |
| Value 9 | Tropical or sub-tropical grassland |
| Value 10 | Temperate or sub-polar grassland |
| Value 11 | Sub-polar or polar shrubland-lichen-moss |
| Value 12 | Sub-polar or polar grassland-lichen-moss |
| Value 13 | Sub-polar or polar barren-lichen-moss |
| Value 14 | Wetland |
| Value 15 | Cropland |
| Value 16 | Barren lands |
| Value 17 | Urban |
| Value 18 | Water |
| Value 19 | Snow and ice |

### RDRSv2_met_forcing

This directory contains the RDRS v2 (Gasset et al., 2021) NetCDF forcing files at a daily timestep (used by most participating models) for the NCRB (Jan-2000 to Dec-2017). The RDRS-v2 data are available over North America at an approximately 10 km grid resolution and an hourly time scale from Jan-2000 to Dec-2017. RDRS-v2 data were downloaded from CaSPAr (www.caspar-data.ca). The original RDRS-v2 data are in UTC time and therefore, they were shifted -5h to be in local time and aggregated to daily values. The following table gives a brief description of the different variables of RDRS-v2 (RDRS-v2 variable name is included in the NetCDF file name):

| Variable short name | Variable long name | Unit | Level | File size |
|---|---|---|---|---|
| PR0 | Quantity of precipitation (CaPA) | m | SFC | ~10.5 GB |
| FB | Downward solar flux | W/m2 | SFC | ~3.3 GB |
| FI | Surface incoming infrared flux | W/m2 | SFC | ~6.6 GB |
| HU | Specific humidity | kg/kg | ~40m | ~8.6 GB |
| HR | Relative humidity | Fraction | ~40m | ~10.2 GB |
| PO | Surface pressure | mb | SFC | ~8.5 GB |
| TT | Air temperature | °C | 1.5m | ~9.9 GB |
| UUC | U-component of the wind (along the grid X axis) corrected for rotated grid | kts | 10m | ~16.6 GB |
| VVC | V-component of the wind (along the grid Y axis) corrected for rotated grid | kts | 10m | ~16.6 GB |
| UVC | Wind Modulus (derived using UU and VV). Note: To be used as input wind speed field. | kts | 10m | ~14.9 GB |

## 03 verification datasets

Verification datasets were used to evaluate the candidate models. The datasets were cropped to the NCRB and were aggregated or re-produced in NetCDF format for the convenience of the participating modeling teams.

### GLEAM_AET

The AET values from the GLEAM dataset (E variable from v3.5b dataset, mm/day; Martens et al., 2017) were cropped to the NCRB and are provided in a NetCDF format (Jan-2003 to Sep-2017). The GLEAM data are available at a daily time resolution from 2003 to 2017 on a 0.25-degree grid globally and are downloaded from https://www.gleam.eu/.

### ERA5-Land_SWE

SWE values from ERA5-Land dataset (sd variable in the dataset, meters; Balsamo et al., 2015) were cropped to the NCRB boundary for the simulation period (Oct-2000 to Sep-2017) and were aggregated to daily average values to be used for model evaluation. The original ERA5-Land data are in UTC time and therefore, they were shifted -5h to be in local time and aggregated to daily values. The ERA5-Land data are available at an hourly time resolution from 1950 to present at a 0.1-degree grid globally and were downloaded from https://doi.org/10.24381/cds.e2161bac.

## 04 model results

This directory contains the results of the different models. Each model has its own parent directory and multiple child directories for different results. The following directory structure applies to any model. An example from the HYPE model (HYPE.zip) was used below for demonstration purposes.

- *Phase_0*: results for Phase 0 of the Nelson-MIP project
  - *Flow*: streamflow simulation results for the different simulated gauges in csv/xlsx format. The csv/xlsx files have the following name code: gauge_id_model_name_avg_results.csv; for example, 05AA022_HYPE_avg_results.csv is the streamflow simulation results for gauge 05AA022 for the HYPE model. Each simulated gauge has a separate file with the following column header (DATE --> date in yyyy-mm-dd format, QOBS --> observed flows in cms, and QSIM --> simulated flows in cms). The files may contain additional columns, but only DATE, QOBS, and QSIM are of importance (i.e., ignore the additional columns).
- *Phase_1*: results for Phase 1 of the Nelson-MIP project
  - *Flow*: streamflow simulation results for the different simulated gauges in csv/xlsx format. The csv/xlsx files have the same name code and format as Phase 0 above.
  - *Hyd.rar*: AET and SWE simulation results for the simulation resolution of the model (watershed for lumped, sub-watershed or grid for semi-distributed models) in csv (timeseries) and shp (polygons and metrics) format.
    - *ERA5-Land_obs_SWE_to_HYPE_native_res.csv*: ERA5-Land SWE values aggregated at the simulation resolution (sub-watershed, native_resolution) of HYPE.
    - *HYPE_sim_SWE_to_native_res.csv*: HYPE's simulated SWE values at the simulation resolution (sub-watershed, native_resolution).
    - *GLEAM_obs_AET_to_HYPE_native_res.csv*: GLEAM AET values aggregated at the simulation resolution (sub-watershed, native_resolution) of HYPE.
    - *HYPE_sim_AET_to_native_res.csv*: HYPE's simulated AET values at the simulation resolution (sub-watershed, native_resolution).
    - *HYPE_native_res.shp*: ESRI shapefile with the geometry of HYPE's sub-watersheds (as delineated by the HYPE modelling team) with AET and SWE performance metrics (r and KGE)

    The csv files under the *Hyd.rar* directory have the following format: the first column is the date is in yyyy-mm-dd format and the rest of the columns represent the timeseries of the hydrological variable (AET or SWE based on file) for each simulation resolution unit (sub-watershed in case of HYPE) (the column has the SubID field identified from the *HYPE_native_res.shp* as the column header).

The *00 Raw KGE metrics* parent directory contains the KGE metric for the different phases, simulation periods, and models at the 132 gauges.

## References

- Albers, S., 2017. tidyhydat: Extract and Tidy Canadian Hydrometric Data. The Journal of Open Source Software 2. https://doi.org/10.21105/joss.00511
- Balsamo, G., Albergel, C., Beljaars, A., Boussetta, S., Brun, E., Cloke, H., Dee, D., Dutra, E., Muñoz-Sabater, J., Pappenberger, F., De Rosnay, P., Stockdale, T., Vitart, F., 2015. ERA-Interim/Land: A global land surface reanalysis data set. Hydrol. Earth Syst. Sci. 19, 389–407. https://doi.org/10.5194/hess-19-389-2015
- Cicco, L.A.D., Lorenz, D., Hirsch, R.M., Watkins, W., Johnson, M., 2022. dataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services. U.S. Geological Survey, Reston, VA. https://doi.org/10.5066/P9P4L3GE
- Gasset, N., Fortin, V., Dimitrijevic, M., Carrera, M., Bilodeau, B., Muncaster, R., Gaborit, É., Roy, G., Pentcheva, N., Bulat, M., Wang, X., Pavlovic, R., Lespinas, F., Khedhaouiria, D., Mai, J., 2021. A 10 km North American precipitation and land-surface reanalysis based on the GEM atmospheric model. Hydrol. Earth Syst. Sci. 25, 4917–4945. https://doi.org/10.5194/hess-25-4917-2021
- Latifovic, R., Homer, C., Ressl, R., Pouliot, D., Nazmul Hossain, S., Colditz, R.R., Olthof, I., Giri C., and Victoria, A. 2012. North American Land Change Monitoring System, Remote sensing of land use and land cover: principles and applications, pp 303-324, Taylor & Francis Group, ISBN 978-1-41200-7074-3
- Martens, B., Miralles, D.G., Lievens, H., Van Der Schalie, R., De Jeu, R.A.M., Fernández-Prieto, D., Beck, H.E., Dorigo, W.A., Verhoest, N.E.C., 2017. GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. Geosci. Model Dev. 10, 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017
- Shangguan, W., Dai, Y., Duan, Q., Liu, B., Yuan, H., 2014. A global soil data set for earth system modeling. J. Adv. Model. Earth Syst. 6, 249–263. https://doi.org/10.1002/2013MS000293
- Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P.D., Allen, G.H., Pavelsky, T.M., 2019. MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset. Water Resour. Res. 55, 5053–5073. https://doi.org/10.1029/2019WR024873