Global mapping of groundwater-dependent ecosystems in drylands

Technical documentation for:

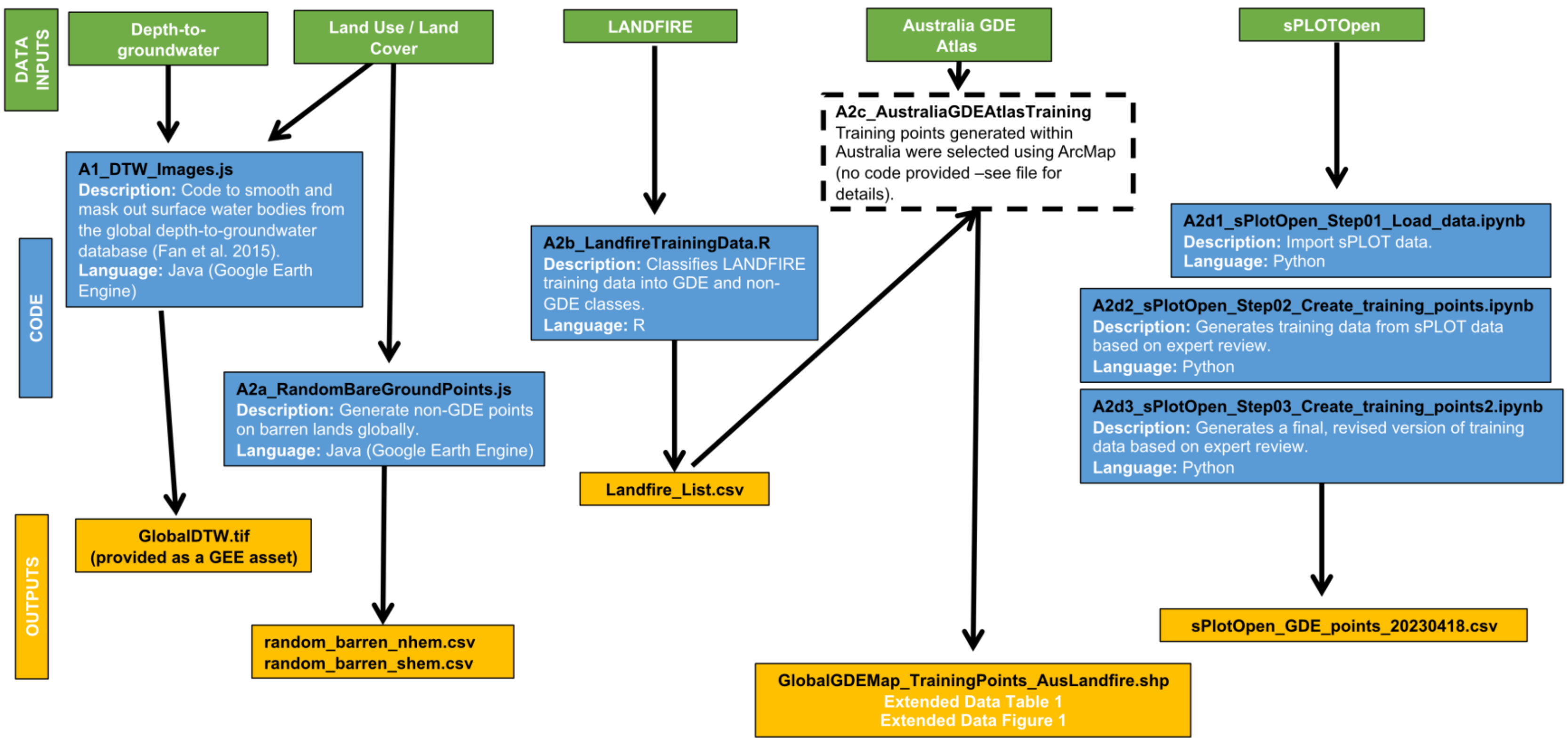
Rohde, M.M., C.M. Albano, X. Huggins, K.R. Klausmeyer, C. Morton, A. Sharman, E. Zaveri, L. Saito, Z. Freed, J.K. Howard, N. Job, H. Richter, K. Toderich, A. Rodella, T. Gleeson, J. Huntington, H.A. Chandanpurkar, A.J. Purdy, J.S. Famiglietti, M.B. Singer, D.A. Roberts, K. Caylor, J.C. Stella. Groundwater-dependent ecosystem map exposes global dryland protection needs. *Nature* (2024).
DOI: 10.1038/s41586-024-07702-8

Prepared by Melissa M. Rohde and Xander Huggins
melissa@RohdeEnvironmental.com
xanderhuggins@uvic.ca

This technical documentation outlines the data and code used to map and analyze groundwater-dependent ecosystems across global drylands, as well as produce final data outputs. Each of the four sections contain a flow chart illustrating what data inputs (green boxes) were used to execute code (blue boxes), and the resulting outputs (gold boxes). Individual scripts are described in tables below each flow chart.

Section A **GDE Model preparation**	**Section B** **GDE Random Forest model**	**Section C** **Post hoc analysis & result reporting**	**Section D** **Data access**
• Preprocesses input data for random forest model. • Generates training point dataset for the random forest model. Output from these scripts feeds into section B.	Random forest (RF) model development including: • Assigning predictor variables to each training point for model ingestion. • Hyperparameter tuning. • Running RF model. • Model validation: Cross Validation tests. • Model validation: distribution plots of training data. Generates core output of this study: **Global 30m GDE classification map and GDE probability map.**	• Prepares groundwater storage and protected area data for comparison to GDE map. • Assesses GDE distributions against pastoral lands, groundwater storage trends, protection status datasets. • Derives all quantitative results. • Prepares aggregated data products for deposition (see Section D).	• Description of all deposited data products: - 1 arcsec (~30 m), - 30 arcsec (~1 km), - 5 arcmin (~10 km), - 30 arcmin (~50 km). • Data access URLs.
PAGE 2 [go to section]	**PAGE 5** [go to section]	**PAGE 20** [go to section]	**PAGE 25** [go to section]

Section A
GDE Model Preparation



DATA INPUTS: Data inputs are provided in the Zenodo repo. The original data inputs used are based on publicly available datasets that need to be directly downloaded from the weblinks included within the code.
CODE: All code was performed using the Google Earth Engine code editor (Java), R version 4.3.1, or Python 3.9.15
OUTPUT: Output data files are provided either in the Zenodo repository and/or as a Google Earth Engine (GEE) asset.

A1 – Prepare masking data

Script name	A1_DTG_Images.js
Description	Smooths and masks out surface water bodies from the global depth-to-groundwater database (Fan et al. 2017).
Data inputs	Fan et al. 2017 (subset by continent; Google Earth Engine asset) ESRI 10 m Land Use Land Cover data (Google Earth Engine asset)
Intermediary data outputs	GlobalDTG_1.tif, GlobalDTG_2.tif (provided as a Google Earth Engine assets)
Authors	Christine Albano and Melissa M. Rohde

A2 – Generate Training Data

Script name	A2a_RandomBareGroundPoints.js
Description	Generates 10,000 non-GDE points on barren lands globally (5000 points in each hemisphere).
Data inputs	ESRI 10 m Land Use Land Cover data (Google Earth Engine asset)
Intermediary data outputs	random_barren_nhem.csv random_barren_shem.csv
Authors	Christine Albano and Melissa M. Rohde

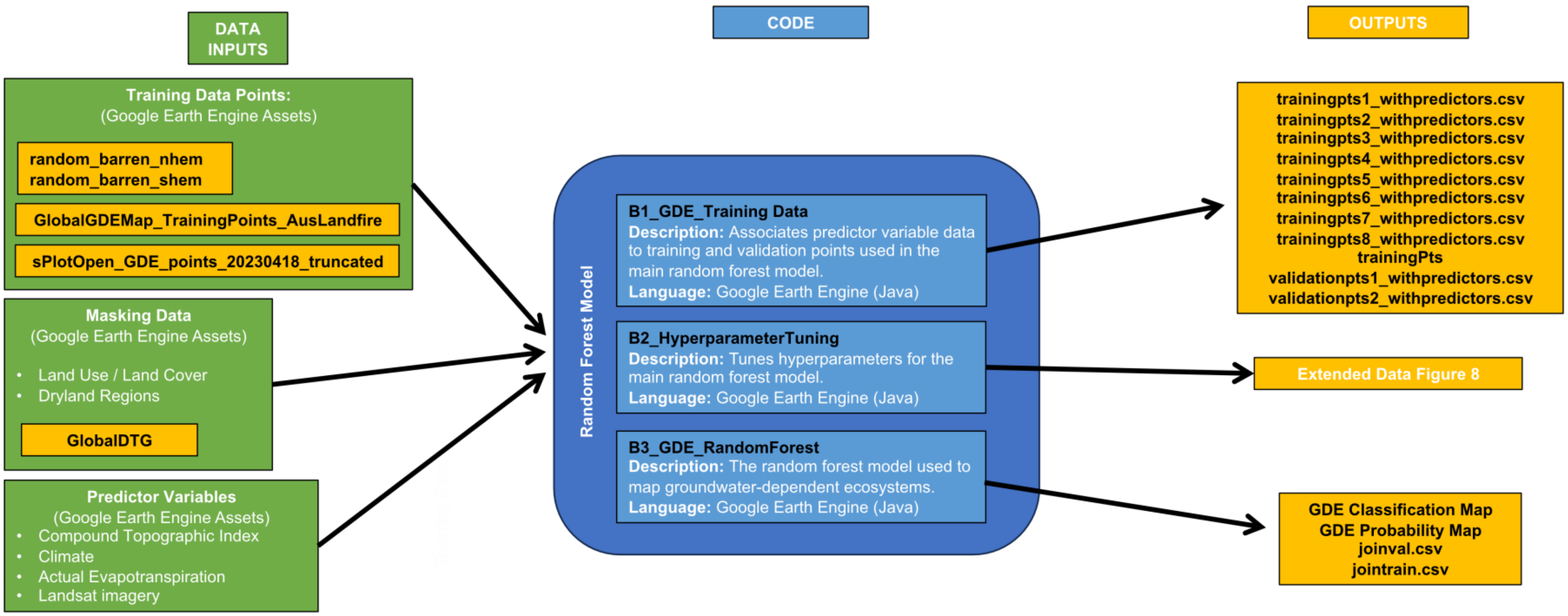
Script name	A2b_LandfireTrainingData.R
Description	Classifies LANDFIRE training data into GDE and non-GDE classes.
Data inputs	vegcamp_GDE_review_20161107.csv: This file contains GDE certainty from California based on Klausmeyer et al. 2018. Landfire_TrainingData.csv: This file contains GDE confirmation based on published data and expert review (see Supplementary Table 4). tbl_DomSp_Unique_NW_SW.csv: This file contains a list of all the unique dominant species names in the LANDFIRE dataset.
R packages used	*dplyr* (Wickham et al., 2023).
Intermediary data outputs	Landfire_List.csv: This file was used to extract training data from the original LANDFIRE reference data in ArcGIS. GDE classification for individual states is as follows: 0=N/A, 1=GDE, 2 = non-GDE. GDE certainty ranges from 1-3, with 1 being the highest, and 3 being the lowest.

Script name	A2c_AustraliaGDEAtlasTraining.rtf
Description	Training points generated within Australia were selected using ArcMap (no code provided). These are the steps for processing the Australian GDE training points.
Data inputs	GDE_Atlas_Aquatic_GDEs.gdb (*n* = 1,107,524 features) GDE_Atlas_Terrestrial_GDEs.gdb (*n* = 7,747,955 features) Both geodatabases are not provided in this data repository. Contact: water@bom.gov.au at the Australian Bureau of Meteorology for bulk download information.
Intermediary data outputs	GlobalGDEMap_TrainingPoints_AusLandfire.shp
Authors	Melissa M. Rohde

Script name	A2d1_sPlotOpen_Step01_Load_data.ipynb A2d2_sPlotOpen_Step02_Create_training_points.ipynb A2d3_sPlotOpen_Step03_Create_training_points2.ipynb
Description	The first python code file includes the steps to import the sPLOT Open dataset and summarize all of the species found in each continent. This table was then shared with expert reviewers to classify each species/continent combination as GDE or not GDE. The second code file takes the results of the expert review and links it back to the sPLOT Open dataset to create the first version of the training points for the machine learning model. The third code file imports a revised version of the expert review file and creates the final set of training points.
Data inputs	sPlotOpen_GDE_review_20230407.csv sPlotOpen_GDE_review_20230407_forToderichReview_Kristina Last_KK.xlsx Both of the above are developed from the sPlotOpen dataset (Sabatini et al., 2021).
Python modules used	*Pandas* (Pandas development team, 2020; McKinney, 2010) *NumPy* (Harris et al., 2020). *FuzzyWuzzy* (Cohen, 2020).
Intermediary data outputs	sPlotOpen_GDE_points_20230418_truncated.csv sPlotOpen_FinalList_20230418.csv sPLOTOpen_GDE_points_20230418.csv
Authors	Kirk Klausmeyer

Section B
GDE Random Forest Model

SUBSECTION:
MODEL DEVELOPMENT



DATA INPUTS: Data inputs are provided as Google Earth Engine assets within the code.
CODE: All code was performed using the Google Earth Engine code editor (Java).
OUTPUT: Output data files (black text) are provided either in the Zenodo repository and/or as a Google Earth Engine asset, or those in white text are provided in the published paper.

B1 – Export Training Data for main model

Script name	B1_GDE_TrainingData.js
Description	Associates predictor variable data to training and validation points used in the B3_GDE_RandomForest mapping model.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data Points: ● GlobalGDEMap_TrainingPoints_AusLandfire ● random_barren_nhem ● random_barren_shem ● sPlotOpen_GDE_points_20230418_truncated
Intermediary data outputs	trainingpts1_withpredictors.csv trainingpts2_withpredictors.csv trainingpts3_withpredictors.csv trainingpts4_withpredictors.csv trainingpts5_withpredictors.csv trainingpts6_withpredictors.csv trainingpts7_withpredictors.csv trainingpts8_withpredictors.csv trainingPts (Google Earth Engine asset) validationpts1_withpredictors.csv validationpts2_withpredictors.csv
Authors	Melissa M. Rohde

B2 - Hyperparameter Tuning

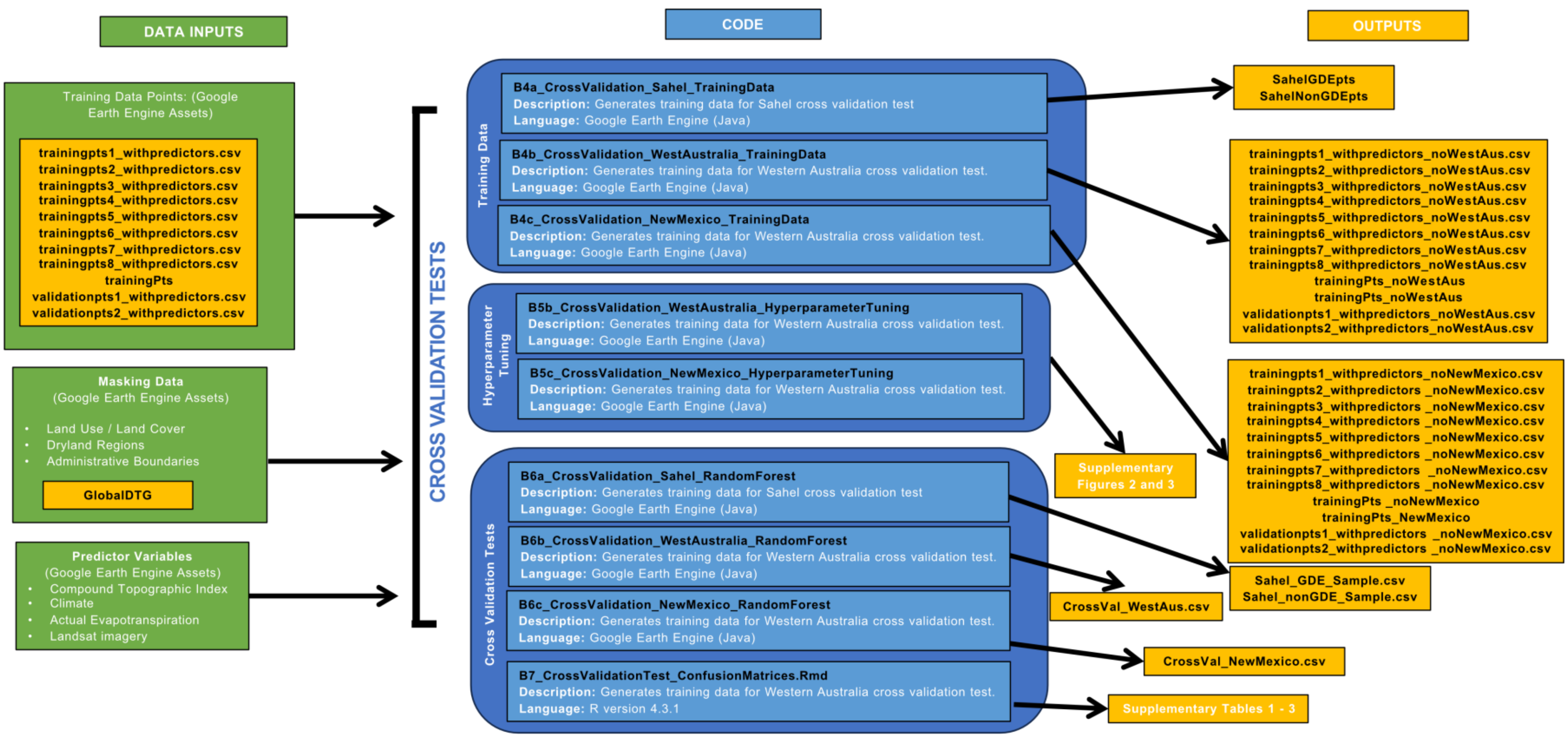
Script name	B2_HyperparameterTuning.js
Description	Retrieves optimal parameters in the random forest model to achieve highest accuracy for GDE classification.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data: ● trainingpts1_withpredictors ● trainingpts2_withpredictors ● trainingpts3_withpredictors ● trainingpts4_withpredictors ● trainingpts5_withpredictors ● trainingpts6_withpredictors ● trainingpts7_withpredictors ● trainingpts8_withpredictors ● validationpts1_withpredictors ● validationpts2_withpredictors
Intermediary data outputs	Accuracy tables and charts for *numberOfTrees, variablesPerSplit, minLeafPopulation, bagFraction,* and *maxNodes* parameters. Output is compiled in Extended Data Figure 8.
Authors	Melissa M. Rohde

B3 - Classify GDEs using the Random Forest model

Script name	B3_GDE_RandomForest.js
Description	The random forest model used to map groundwater-dependent ecosystems globally. Tuned hyperparameters from B2 (above) are included in this model run.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data: ● trainingpts1_withpredictors ● trainingpts2_withpredictors ● trainingpts3_withpredictors ● trainingpts4_withpredictors ● trainingpts5_withpredictors ● trainingpts6_withpredictors ● trainingpts7_withpredictors ● trainingpts8_withpredictors ● validationpts1_withpredictors ● validationpts2_withpredictors
Intermediary data outputs	GDE classification map (Google Earth Engine asset) GDE probability map (Google Earth Engine asset) joinval.csv jointrain.csv
Authors	Christine Albano and Melissa M. Rohde

Section B:
GDE Random Forest Model

SUBSECTION:
MODEL VALIDATION - CROSS VALIDATION



DATA INPUTS: Data inputs are provided as Google Earth Engine assets within the code.
CODE: All code was performed using the Google Earth Engine code editor (Java) or R version 4.3.1.
OUTPUT: Output data files (black text) are provided either in the Zenodo repository and/or as a Google Earth Engine asset, or those in white text are provided in the published paper.

B4 - Generate training data for regional cross validation tests

Script name	B4a_CrossValidation_Sahel_TrainingData.js
Description	Generates training data points using GDE location data from the World Bank that were extracted from peer-review literature sources, and non-GDE points from the ESRI land use and land cover bare ground layer.
Data inputs	GDE lines and point data (Rodella et al., 2023) ESRI 10 m Land Use Land Cover data (Google Earth Engine asset)
Intermediary data outputs	SahelGDEpts (Google Earth Engine asset) SahelNonGDEpts (Google Earth Engine asset)
Authors	Melissa M. Rohde

Script name	B4b_CrossValidation_WestAustralia_TrainingData.js
Description	This Google Earth Engine code generates training data points for the Western Australia cross validation test.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data Points: ● Administrative boundaries (FAO/GAUL/2015/level1) ● GlobalGDEMap_TrainingPoints_AusLandfire ● random_barren_nhem ● random_barren_shem ● sPlotOpen_GDE_points_20230418_truncated
Intermediary data outputs	trainingpts1_withpredictors_noWestAus.csv trainingpts2_withpredictors_noWestAus.csv trainingpts3_withpredictors_noWestAus.csv trainingpts4_withpredictors_noWestAus.csv trainingpts5_withpredictors_noWestAus.csv trainingpts6_withpredictors_noWestAus.csv trainingpts7_withpredictors_noWestAus.csv trainingpts8_withpredictors_noWestAus.csv trainingPts_noWestAus (Google Earth Engine asset) trainingPts_WestAus (Google Earth Engine asset)

	validationpts1_withpredictors_noWestAus.csv validationpts2_withpredictors_noWestAus.csv
Authors	Melissa M. Rohde

Script name	B4c_CrossValidation_NewMexico_TrainingData.js
Description	Generates training data points for the New Mexico cross validation test.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data Points: ● Administrative boundaries (FAO/GAUL/2015/level1) ● GlobalGDEMap_TrainingPoints_AusLandfire ● random_barren_nhem ● random_barren_shem ● sPlotOpen_GDE_points_20230418_truncated
Intermediary data outputs	trainingpts1_withpredictors_noNewMexico.csv trainingpts2_withpredictors_noNewMexico.csv trainingpts3_withpredictors_noNewMexico.csv trainingpts4_withpredictors_noNewMexico.csv trainingpts5_withpredictors_noNewMexico.csv trainingpts6_withpredictors_noNewMexico.csv trainingpts7_withpredictors_noNewMexico.csv trainingpts8_withpredictors_noNewMexico.csv trainingPts_noNewMexico (Google Earth Engine asset) trainingPts_NewMexico (Google Earth Engine asset) validationpts1_withpredictors_noNewMexico.csv validationpts2_withpredictors_noNewMexico.csv
Authors	Melissa M. Rohde

B5 - Model Validation: Cross Validation Tests. Hyperparameter tuning for regional cross validation tests

Note: Hyperparameter tuning for the Sahel cross validation test utilizes the tuned parameter outputs from B2_HyperparameterTuning.js.

Script name	B5b_CrossValidation_WestAustralia_HyperparameterTuning.js
Description	Tunes parameters in the random forest model for the West Australia cross validation test to retrieve optimal parameters to achieve highest accuracy.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data: ● trainingpts1_withpredictors_noWestAus ● trainingpts2_withpredictors_noWestAus ● trainingpts3_withpredictors_noWestAus ● trainingpts4_withpredictors_noWestAus ● trainingpts5_withpredictors_noWestAus ● trainingpts6_withpredictors_noWestAus ● trainingpts7_withpredictors_noWestAus ● trainingpts8_withpredictors_noWestAus ● trainingPts_noWestAus ● trainingPts_WestAus ● validationpts1_withpredictors_noWestAus ● validationpts2_withpredictors_noWestAus
Intermediary data outputs	Accuracy tables and charts for *numberOfTrees, variablesPerSplit, minLeafPopulation, bagFraction,* and *maxNodes* parameters.
Authors	Melissa M. Rohde

Script name	B5c_CrossValidation_NewMexico_HyperparameterTuning.js
Description	Tunes parameters in the random forest model for the New Mexico cross validation test to retrieve optimal parameters to achieve highest accuracy.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG

	● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data: ● trainingpts1_withpredictors_noNewMexico ● trainingpts2_withpredictors_noNewMexico ● trainingpts3_withpredictors_noNewMexico ● trainingpts4_withpredictors_noNewMexico ● trainingpts5_withpredictors_noNewMexico ● trainingpts6_withpredictors_noNewMexico ● trainingpts7_withpredictors_noNewMexico ● trainingpts8_withpredictors_noNewMexico ● trainingPts_noNewMexico ● trainingPts_NewMexico ● validationpts1_withpredictors_noNewMexico ● validationpts2_withpredictors_noNewMexico
Intermediary data outputs	Accuracy tables and charts for *numberOfTrees, variablesPerSplit, minLeafPopulation, bagFraction,* and *maxNodes* parameters.
Authors	Melissa M. Rohde

B6 - Model Validation: Cross Validation Tests. Perform regional cross validation tests with tuned hyperparameters

Script name	B6a_CrossValidation_Sahel_RandomForest.js
Description	Samples the main global GDE classification output using the Sahel training points.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data: ● trainingpts1_withpredictors ● trainingpts2_withpredictors ● trainingpts3_withpredictors ● trainingpts4_withpredictors ● trainingpts5_withpredictors ● trainingpts6_withpredictors

	• trainingpts7_withpredictors • trainingpts8_withpredictors • validationpts1_withpredictors • validationpts2_withpredictors • trainingPts • SahelGDEpts • SahelNonGDEpts
Intermediary data outputs	Sahel_GDE_Sample.csv Sahel_nonGDE_Sample.csv
Authors	Melissa M. Rohde

Script name	B6b_CrossValidation_WestAustralia_RandomForest.js
Description	Cross validation test for Western Australia.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: • GlobalDTG • ESRI 10 m Land Use Land Cover data • Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: • Compound Topographic Index (TopConvIndex_Global_KGclims) • Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) • Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) • Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data • trainingpts1_withpredictors_noWestAus • trainingpts2_withpredictors_noWestAus • trainingpts3_withpredictors_noWestAus • trainingpts4_withpredictors_noWestAus • trainingpts5_withpredictors_noWestAus • trainingpts6_withpredictors_noWestAus • trainingpts7_withpredictors_noWestAus • trainingpts8_withpredictors_noWestAus • trainingPts_noWestAus • trainingPts_WestAus • validationpts1_withpredictors_noWestAus • validationpts2_withpredictors_noWestAus
Intermediary data outputs	CrossVal_WestAus.csv
Authors	Melissa M. Rohde

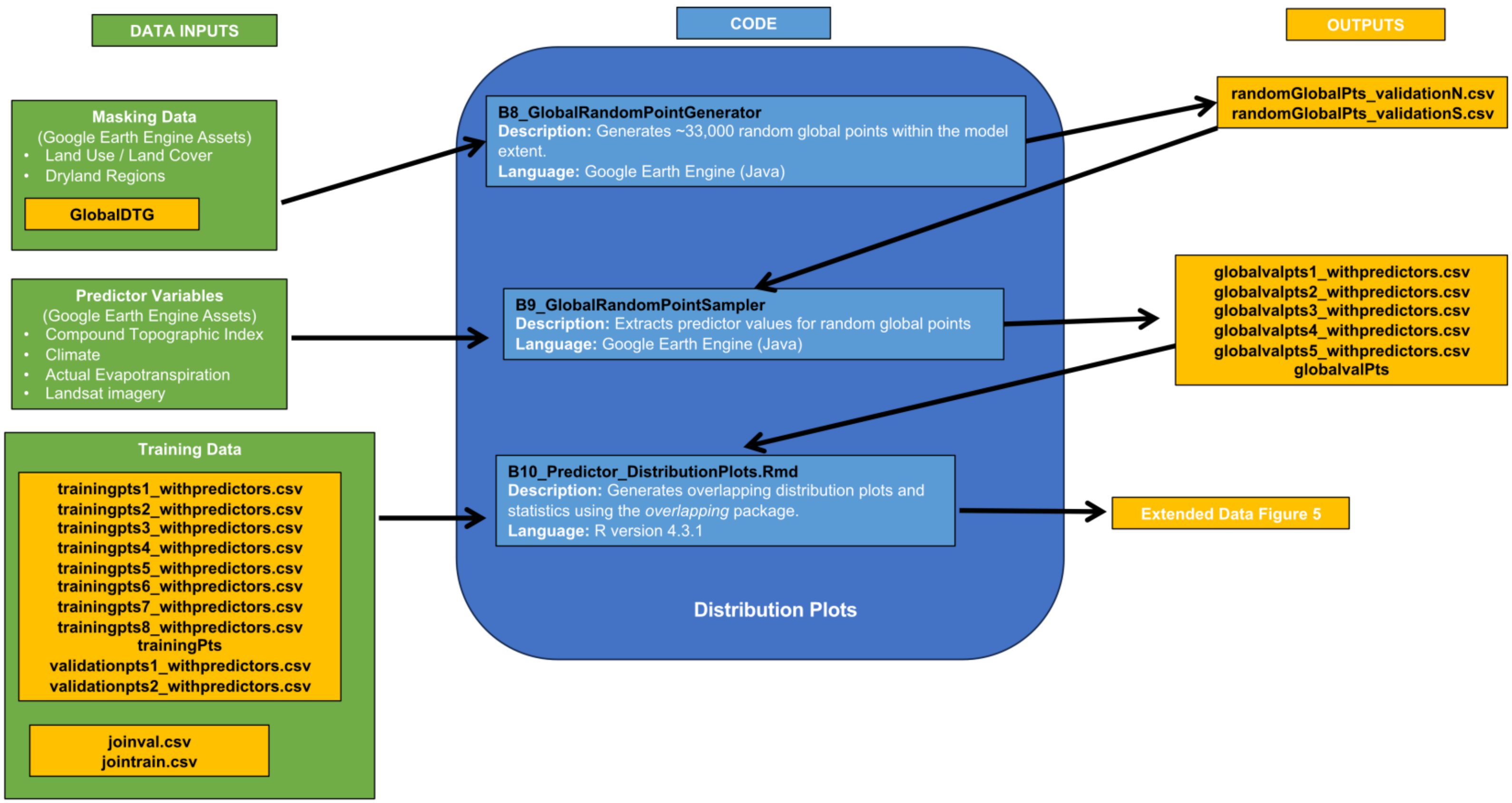
Script name	B6c_CrossValidation_NewMexico_HyperparameterTuning.js
Description	Cross validation test for New Mexico.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) Training Data: ● trainingpts1_withpredictors_noNewMexico ● trainingpts2_withpredictors_noNewMexico ● trainingpts3_withpredictors_noNewMexico ● trainingpts4_withpredictors_noNewMexico ● trainingpts5_withpredictors_noNewMexico ● trainingpts6_withpredictors_noNewMexico ● trainingpts7_withpredictors_noNewMexico ● trainingpts8_withpredictors_noNewMexico ● trainingPts_noNewMexico ● trainingPts_noNewMexico ● validationpts1_withpredictors_noNewMexico ● validationpts2_withpredictors_noNewMexico
Intermediary data outputs	CrossVal_NewMexico.csv
Authors	Melissa M. Rohde

B7 - Model Validation: Cross Validation Tests. Generate confusion matrices for cross validation tests

Script name	B7_CrossValidationTests_ConfusionMatrices.Rmd
Description	Generate confusion matrices for the Sahel, Western Australia, and New Mexico cross validation tests.
Data inputs	Sahel_GDE_Sample.csv Sahel_nonGDE_Sample.csv CrossVal_WestAus.csv CrossVal_NewMexico.csv
R packages used	*overlapping* (Pastore et al., 2022). *gridExtra* (Auguie & Antonov, 2017). *tidyverse* (Wickham et al., 2019)
Intermediary data outputs	Confusion matrix for each cross-validation test (Supplementary Tables 1-3).

Section B:
GDE Random Forest Model

SUBSECTION:
MODEL VALIDATION – DISTRIBUTION PLOTTING



DATA INPUTS: Data inputs are provided as Google Earth Engine assets within the code or in the Zenodo repository.
CODE: All code was performed using the Google Earth Engine code editor (Java) or R version 4.3.1.
OUTPUT: Output data files (black text) are provided either in the Zenodo repository and/or as a Google Earth Engine asset, or those in white text are provided in the published paper.

B8 - Generate Random global points within model extent to compare distribution of variables within model extent and training data used in model.

Script name	B8_GlobalRandomPointGenerator.js
Description	This Google Earth Engine code generates 125,000 random points, retaining only those that fall within the model extent (n~33,000 global points).
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083)
Intermediary data outputs	randomGlobalPts_validationN.csv randomGlobalPts_validationS.csv
Authors	Melissa M. Rohde

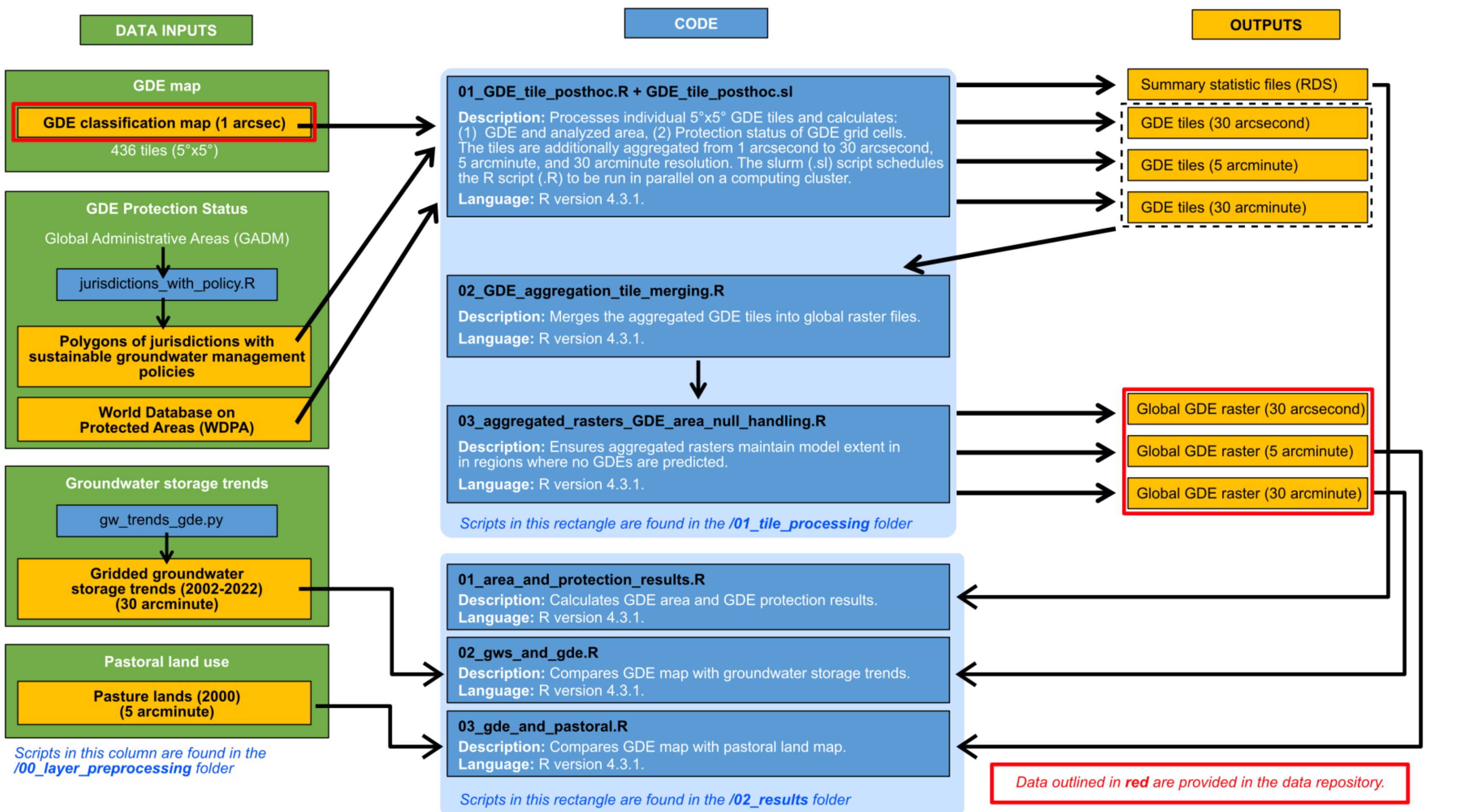
B9 - Extract predictor variable values for each random global point from model extent

Script name	B9_GlobalRandomPointSampler.js
Description	This Google Earth Engine code extracts predictor values at randomly generated points within the model extent to compare distributions with those of training points in the model.
Data inputs	All data inputs are uploaded and accessible as Google Earth Engine assets. Links are provided in the code. Masking Data: ● GlobalDTG ● ESRI 10 m Land Use Land Cover data ● Dryland Regions (Beck_KG_V1_present_0p0083) Predictor Variables: ● Compound Topographic Index (TopConvIndex_Global_KGclims) ● Climate - precipitation and potential evapotranspiration (IDAHO_EPSCOR/TERRACLIMATE) ● Actual Evapotranspiration (CAS/IGSNRR/PML/V2_v017) ● Landsat imagery (LANDSAT/LC08/C02/T1_L2) ● randomGlobalPts_validationN ● randomGlobalPts_validationS
Intermediary data outputs	globalvalpts1_withpredictors.csv globalvalpts2_withpredictors.csv globalvalpts3_withpredictors.csv globalvalpts4_withpredictors.csv globalvalpts5_withpredictors.csv globalvalPts.csv
Authors	Melissa M. Rohde

B10 - Generate distribution plots of predictor values for the global and model training points.

Script name	B10_Predictor_DistributionPlots.Rmd
Description	Generates overlapping distribution plots and statistic.
Data inputs	globalvalpts1_withpredictors globalvalpts2_withpredictors globalvalpts3_withpredictors globalvalpts4_withpredictors globalvalpts5_withpredictors globalvalPts trainingpts1_withpredictors trainingpts2_withpredictors trainingpts3_withpredictors trainingpts4_withpredictors trainingpts5_withpredictors trainingpts6_withpredictors trainingpts7_withpredictors trainingpts8_withpredictors trainingPts validationpts1_withpredictors validationpts2_withpredictors joinval.csv jointrain.csv
R packages used	*overlapping* (Pastore et al., 2022). *gridExtra* (Auguie & Antonov, 2017). *tidyverse* (Wickham et al., 2019)
Intermediary data outputs	Distribution plots and statistics (Extended Data Figure 5).
Authors	Melissa M. Rohde

Section C
Post hoc analysis and result reporting



Section C
Post hoc analysis and result reporting

SUBSECTION:
POSTHOC LAYER PREPROCESSING

Script name	00_layer_preprocesing/**gw_trends_gde.py**
Description	Derives GRACE-based gridded groundwater storage trends at 30 arcminute resolution using GLDAS-2.1 output from both Noah and VIC land surface models. The script provided shows the workflow using Noah model output, however the same is also performed with VIC model output.
Python packages used	*NumPy* (Harris et al., 2020). *xarray* (Hoyer & Harmann, 2017). *mvstats* (Chandanpurkar, 2018).
Intermediary data outputs	gldas_2.1_vic_local_trends.nc gldas_2.1_noah_local_trends.nc
Authors	Hrishikesh A. Chandanpurkar

Script name	00_layer_preprocesing/**jurisdictions_with_policy.R**
Description	Generates a vector file representing jurisdictions with sustainable groundwater management policies.
Python packages used	*terra* (Hijmans, 2023).
Intermediary data outputs	all_juris_with_GDE_protection_policies.sqlite
Authors	Xander Huggins

Section C:
Post hoc analysis and result reporting

SUBSECTION:
TILE PROCESSING

Script name	01_tile_processing/**01_GDE_tile_posthoc.R**
Description	For each 5-degree GDE tile at 1 arsecond resolution, this script: 1) Calculates grid cell area. 2) Computes GDE area, analyzed area, and grid cell area. 3) Rasterizes the WDPA and extent of jurisdictions with sustainable groundwater management policies at 1 arcsecond resolution. 4) Zonally summarizes GDE area within different forms of protection. 5) Computes GDE area and analyzed area per continent. 6) Aggregates GDE area, analyzed area, and grid cell area to the resolutions of: 30 arcseconds, 5 arcminutes, and 30 arcminutes. 7) Computes (i) GDE area to analyzed area, (ii) GDE area to grid cell area, and (iii) analyzed area to grid cell area fractions.
R packages used	*terra* (Hijmans, 2023). *raster* (Hijmans, 2023). *rasterDT* (O'Brien, 2022). *readr* (Wickham et al., 2023).
Intermediary data outputs	Zonal summary statistic file per tile. Aggregated GDE tiles at 30 arcsecond, 5 arcminute, and 30 arcminute resolution.
Authors	Xander Huggins

Script name	01_tile_processing/**GDE_tile_posthoc.sl**
Description	Slurm script to run GDE tile post hoc analyses on cluster. The R script scheduled through this script is provided above (01_GDE_tile_posthoc.R).
Cluster acknowledgement	Computing resources were provided by the Digital Research Alliance of Canada (https://alliancecan.ca/).
Authors	Xander Huggins

Script name	01_tile_processing/**02_GDE_aggregation_tile_merging.R**
Description	Merges the exported aggregated (i.e., 30s, 5m, 30m) tiles into global (single) rasters.
R Packages	*terra* (Hijmans, 2023).
Authors	Xander Huggins

Script name	01_tile_processing/**03_aggregated_rasters_GDE_area_null_handling.R**
Description	Ensures that grid cells with GDE area = 0 and analyzed area > 0 are reported with GDE areas and GDE area fractions of 0 rather than NA.
Data outputs	Global aggregated GDE rasters at 30 arcsecond, 5 arcminute, and 30 arcminute resolution for deposition on the data repository.
R Packages	*terra* (Hijmans, 2023).
Authors	Xander Huggins

Section C:
Post hoc analysis and result reporting

SUBSECTION:
RESULT REPORTING

Script name	02_results/ **01_area_and_protection_results.R**
Description	Reports base GDE results such as GDE area and analyzed area per continent, and GDE protection stats.
R Packages	*terra* (Hijmans, 2023).
Authors	Xander Huggins

Script name	02_results/ **02_gws_and_gde.R**
Description	Compares GDE area density to groundwater storage trends at 30 arcminute resolution. Develops the bi-variate map as plotted in Figure 2. Calculates area-weighted groundwater storage trends and GDE area density for select freshwater ecoregions of the world.
R Packages	*terra* (Hijmans, 2023). *tidyverse* (Wickham et al., 2019).
Authors	Xander Huggins

Script name	02_results/ **03_ gde_and_pastoral.R**
Description	Develops bi-variate map as plotted in Extended Data Figure 7, comparing GDE area density with pasture land area density. Calculates percentage of GDE area that lies within regions with at least 25% pasture land area density.
R Packages	*terra* (Hijmans, 2023). *tidyverse* (Wickham et al., 2019).
Authors	Xander Huggins

Section D
Data deposit and access

Global GDE data are deposited and openly accessible at four resolutions. All raster layers are referenced to the World Geodetic System 1984 (WGS84) coordinate reference system.

Core GDE data:

Dataset 1: 1 arcsecond GDE data (~30 m grids at the equator)

At this base resolution, two raster layers are provided:

Layer: ***classification***
Values:
0: Pixel is outside of model domain
1: Pixel is likely a GDE
2: Pixel is not likely a GDE

Layer: ***probability***
Values:
[0-100]: Likelihood pixel is a GDE (100) or non-GDE (0)

These layers are provided across 436 tiles, each with a 5° x 5° extent. For example, the tile named "n45w105.tif" has an extent of: xmin: -105, xmax: -100, ymin: 45, ymax: 50.

Access:
These tiles can be accessed on the data repository (https://doi.org/10.5281/zenodo.11062894) by navigating to the folder named **D_Data_GDEtiles_5deg**.

Aggregated GDE data:

Dataset 2: Aggregated GDE data at 30 arcsecond resolution (~1 km grids at the equator)
Dataset 3: Aggregated GDE data at 5 arcminute resolution (~10 km grids at the equator)
Dataset 4: Aggregated GDE data at 30 arcminute resolution (~50 km grids at the equator)

At these aggregated resolutions, five raster layers are provided. To manage file size of global rasters, all layers are provided in the INT4U datatype (i.e., capable of storing integers from 0 to 2^{32}). Operating within this upper constraint, fraction layers (normally ranging [0-1]) are multiplied by 10^8 to retain the greatest precision possible. Area layers, across all resolutions, do not have grid cell values that exceed the upper limit of the datatype and thus are not modified other than providing area values that are rounded to the nearest square meter. All aggregated grid cells with no analyzed area are set to `NA`.

Layer: ***GDE_sqm***
Values: Range varies per dataset.
GDE area (m^2) within grid cell.

Layer: ***GDE_frac_AA***
Values: [0-10^8]
Fraction of analyzed area (i.e., area within the model domain) that is a GDE. This fraction is multiplied by 10^8 and saved as an integer. To convert this layer to the range [0-1], divide by 10^8.

Layer: ***GDE_frac_GA***
Values: [0-10^8]
Fraction of grid cell area that is a GDE. This fraction is multiplied by 10^8 and saved as an integer. To convert this layer to the range [0-1], divide by 10^8.

Layer: ***AA_sqm***
Values: Range varies per dataset.
Analyzed area (m^2) within grid cell, rounded to nearest square meter.

Layer: ***AA_frac_GA***
Values: [0-1e8]
Fraction of grid cell area that is analyzed (i.e., within the model domain). This fraction is multiplied by 10^8 and saved as an integer. To convert this layer to the range [0-1], divide by 10^8.

Access:
These datasets can be accessed on the data repository (https://doi.org/10.5281/zenodo.11062894) by navigating to the folder named **D_Data_GDE_AggregatedLayers**.

License agreement

Data citation

Rohde, M.M., C.M. Albano, X. Huggins, K.R. Klausmeyer, C. Morton, A. Sharman, E. Zaveri, L. Saito, Z. Freed, J.K. Howard, N. Job, H. Richter, K. Toderich, A. Rodella, T. Gleeson, J. Huntington, H.A. Chandanpurkar, A.J. Purdy, J.S. Famiglietti, M.B. Singer, D.A. Roberts, K. Caylor, J.C. Stella. Groundwater-dependent ecosystem map exposes global dryland protection needs. *Nature* (2024). DOI: 10.1038/s41586-024-07702-8

References

Auguie, B., & Antonov, A. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://cran.r-project.org/package=gridExtra (2017).

Chandanpurkar, H.A. mvstats: Vectorized multivariate statistical functions for analyzing multi-dimensional earth system data. Python package version 0.1.0. https://mvstats.readthedocs.io (2018).

Cohen, A. fuzzywuzzy: Fuzzy string matching in python. Python package version 0.18.0. https://pypi.org/project/fuzzywuzzy (2020).

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B. & Otero-Casal, C. Hydrologic regulation of plant rooting depth. *Proc National Acad Sci* **114**, 10572–10577 (2017).

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. *Nature* 585, 357–362 (2020).

Hijmans, RJ.. raster: Geographic Data Analysis and Modeling. R package version 3.6-20, https://cran.r-project.org/package=raster (2023)

Hijmans, R. J. terra: Spatial Data Analysis. R package version 1.7-37 https://CRAN.R-project.org/package=terra (2023).

Hoyer, S. & Hamman, J. xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software* 5(1), 1-10. (2017).

Klausmeyer K., J. Howard, T. Keeler-Wolf, K. Davis-Fadtke, R. Hull, A. Lyons. Mapping Indicators of Groundwater Dependent Ecosystems in California: Methods Report. San Francisco, California. (2018).

McKinney, W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference,* 56-61 2010).

O'Brien, J. rasterDT: Fast Raster Summary and Manipulation. R package version 0.3.2. https://CRAN.R-project.org/package=rasterDT (2022).

Pandas development team. *pandas*. Python package version 1.5.2. https://pandas.pydata.org/ (2020).

Pastore, M., Di Loro, P.A., Mingione, M., Calcagni, A. overlapping: Estimation of overlapping empirical distributions. R package version 2.1. https://cran.r-project.org/package=overlapping (2022).

Rodella, A.-S., Zaveri, E.D., Bertone, F. The Hidden Wealth of Nations : The Economics of Groundwater in Times of Climate Change – Executive Summary. World Bank. http://documents.worldbank.org/curated/en/099145503202323072/P178601171e7ffac1ea0714b5e187c0122449517b07d (2023).

Sabatini, F.M., Lenoir, J., Bruelheide, H. & the sPlot Consortium. sPlotOpen – An environmentally-balanced, open-access, global dataset of vegetation plots. [Dataset] https://doi.org/10.25829/idiv.3474-40-3292 (2021).

Wickham H., Averick M., Bryan J., *et al*. "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686 (2019).

Wickham, H., François, R., Henry, L., *et al.* dplyr: A Grammar of Data Manipulation. R package version 1.1.2 https://cran.r-project.org/package=dplyr (2023).

Wickham, H., Hester, J., Bryan, J. readr: Read Rectangular Text Data. R package version 2.1.4, https://cran.r-project.org/package=readr (2023).