

USGS Notes for Extracting Entities for the Model Catalog:

Goal:

- Create NER Model to identify which methods are used in each model using text from the abstract or website description if an article doesn't exist.
- *NEW Idea: Instead of trying to classify the language into NERs (difficult for methods) – predict title and use similarities of predicted titles*

Entities:

- **First Attempt:** 3 Entities
 - METHOD: Machine learning or statistical model (NN, SVM, etc.)
 - TASK: e.g. Image classification, segmentation, etc.
 - APPLICATION: e.g. tracking rainfall thresholds for landslides, etc.
- Observations of 3 entities: Methods were a mixture of mathematical, ML, statistical, or geological models. Applications often had long (multi-word) entities, but were often preceded by the word, "provides", the model was a failure – however it could have been due to improper training. Could come back and try this again.
- **Second Attempt:** 1 Methods (or 2 methods GEO and ML)
 - Methods = GEO model or ML/Statistical Model (try separate and together)
- Observations: There are so few ML entities in the text, there are too few to train on. GEO methods becomes complex and became geological entities instead of geological model entities
- **Third Attempt:**
 - GEO Words (see if 'make-gold' is better)
 - MODEL - Words after 'model'
- Observations: I create the GEO model with nouns, but it may be better to have adjectives (watershed model or water balance algorithm – watershed/water balance is an adjective) – searching for parts of speech, not just keywords. On first try, I did not include any of the adjectives, only nouns.
- If I want to train for words before 'model', first need more abstracts, second try without patterns. Patterns probably makes it worse.
- GEO model (6) produced weird results where it recognized some GEO terms, missed some, and others highlighted all words!?? Accuracy was worse than earlier models but Make Gold was better. Need to redo this test. Labeled NER incorrectly and couldn't fix later.
- MODEL model (7) There are not enough words before 'model' to train on.

Other Observations while manually scraping text:

- Tasks and applications may overlap. Difficulty distinguishing between the two for some articles.

- Some texts may not have a method.
- One article had a one-sentence section called “Introduction” as well as a short “Model Overview and System Requirements”, but no Abstract section.
- Other sections to consider (though not consistent): “purpose and scope”, “summary”, “Model Overview and System Requirements”
- May reference models outside of text – Future models, other models ...
- Mostly specific mathematical models for waterflow, watershed, water balance, landslides, etc. (not machine learning methods)
- GEO either has a lot of material to highlight or to try to weed out what are actual GEO models. ML has maybe 1-2 recognizable ML/Statistical Model in 20 texts (most are geological modeling). few examples for ML-method, so couldn’t train well.
- Models built on prior models (are those models the method?)
- Accuracy correct? models this far label just about everything as an entity
- Need to tease out the geologic models.
- When trying GEO alone, I realized that I started to label any clear geology noun (volcano, stream, ocean, reservoir, etc.). These are not necessarily geologic models, but could be useful... the models are much harder to separate see GeoModel Notes and examples below
- GeoModel Notes:
 - digital elevation model, diffuse layer model,
 - 3D method of columns (landslides)
 - Vertical fluid flow (seepage flux)
 - “building a 3D domain”
 - Limit-equilibrium analysis, slope-stability analysis, slope stability
 - Aqueous geochemical calculations
 - (1D) transport calculations
 - Ion aqueous model, Pitzer aqueous model
 - Peng-Robinson equation (for gas solubility)
 - Transport simulations, geochemical simulations, heat transport equations,
 - Flow and transport: aquifer, pollutants, sediment, solute, heat, thermal energy,
 - Population (density, estimation, prediction, etc.)
 - Water-balance model, water balance modeling,
 - Migration/Migration Cycle
- See examples below for help labeling:

METHOD 1

These include transient transport with first- and zero - order decay and linear sorption and also steadystate transport with first- and zero - order decay , or Monod degradation .

METHOD 1

This formulation builds upon previous developments by coupling the atmospheric model to the ocean and wave models , providing one - way grid refinement in the ocean model , one - way grid refinement in the wave model , and coupling on refined levels .

METHOD 1

The matrix - solver options include a generalized - minimum - residual (GMRES) Solver and an Orthomin / stabilized conjugate - gradient (CGSTAB) Solver .

Volcano (volcanic ash)

Earthquakes (subduction zone, ground motion prediction

Flooding/Waves

Migration

Reservoir water balance

Aqueous geochemical model

Precipitation/Slope Stability Analysis

Hydrodynamics, Vertical Fluid Flow

data estimation from isopleth map

solute transport, thermal energy transport, sediment transport

watersheds and redevelopment

water quality

evolution of coastal zone

climate change

- Used 'bad examples' with few GEO-like examples and one with just about everything except for questionable GEO terms as below:

GEO 1

A new version of the computer program 1DTempPro extends the original code to include new capabilities for (1) automated parameter estimation ,(2) layer heterogeneity , and (3) time-varying specific discharge . The code serves as an interface to the U.S. Geological Survey model VS2DH and supports analysis of vertical one-dimensional temperature profiles under saturated flow conditions to assess groundwater / surface-water exchange and estimate hydraulic conductivity for cases where hydraulic head is known .