ReadMe file for Pennsylvania Department of Environmental Protection (PA DEP) 26r Detailed Produced Water Compositions (version 1.0)

Database Name: Pennsylvania Department of Environmental Protection (PA DEP) 26r Detailed Produced Water Compositions (version 1.0)

Citation: Mackey, J., Pfander, I., Yesenchak, R., Gardiner, J., Lackey, G., Kutchko, B., Fritz, A., Able, C., Siefert, N. "Pennsylvania Department of Environmental Protection (PA DEP) 26r Detailed Produced Water Compositions (version 1.0)". December 20, 2024, DOI: 10.18141/2483335

Description:

A database of geochemical compositions of aqueous species in produced water reported to the PA DEP. Samples were collected between mid-2012 to early-2020. Data from publicly-available PA DEP 26r reports were scraped from pdf files and cumulated into tabular spreadsheet format for >1000 produced water streams from Marcellus wells in Pennsylvania. In addition to providing the original values, the NETL NEWTS team has reformatted the dataset to allow sample streams to be easily copied into OLI Studio and Geochemist WorkBench (GWB) software for modeling the geochemistry and the recovery of critical minerals, such as lithium, from these produced water streams.

In addition, a version of the dataset has been included with predictions for some missing values in the original dataset using machine learning techniques within CoDaRT software, a public ML software developed by the Nation Energy Technology Laboratory. We have made the Input into CoDaRT and one example output from CoDaRT available in this dataset.

Acknowledgements: This work was performed in support of the U.S. Department of Energy's Fossil Energy and Carbon Management through the National Energy Technology Laboratory's (NETL) Research & Innovation Center's Produced Water Characterization & Treatment FWP [DE-FECM1610260]. We acknowledge and are thankful for the assistance from the PA DEP in obtaining digital copies of the 26r reports either via email or by visiting field offices.

Date of publishing: 12/20/2024

Geographic coordinate system: unknown

Points of Contact: Nicholas Siefert, nicholas.siefert@netl.doe.gov ; Rachel Yesenchak, rachel.yesenchak@netl.doe.gov

NETL Reviewer: Nicholas Siefert, nicholas.siefert@netl.doe.gov

Resources list for data release:

1. **PA DEP 26r Dataset CSV Files**
    a. **PA_DEP_26r_pdf_scrape.csv** – Raw data scraped from pdf files in "long" format, where each row represents a separate measurement (multiple rows associated each unique sample)
    b. **PA_DEP_26r_processed.csv** – Reformatted pdf scraped data in "wide" format, where each row represents a unique sample. Unit conversions performed where applicable.
2. **OLI Studio/Geochemist's Workbench Files**
    a. **PA_DEP_26r_GWB_template.gss** – Template for inputting transposed data into Geochemist's Workbench
    b. **PA_DEP_26r_GWB.csv** - The full dataset formatted for input into Geochemist's Workbench. Data must be transposed prior to input into software.
    c. **PA_DEP_26r_GWB_1val_min.csv** - Dataset records with at least 1 relevant measured value formatted for input into Geochemist's Workbench. Data must be transposed prior to input into software.
    d. **PA_DEP_26r_GWB_4majors_min.csv** - Dataset records with at least 4 major cations measured, formatted for input into Geochemist's Workbench. Data must be transposed prior to input into software.
    e. **PA_DEP_26r_OLI_template.oad** – Template for inputting transposed data into OLI Studio.
    f. **PA_DEP_26r_OLI.csv** - The full dataset formatted for input into OLI Studio. Data must be transposed prior to input into software.
    g. **PA_DEP_26r_OLI_1val_min.csv** - Dataset records with at least 1 relevant measured value formatted for input into OLI Studio. Data must be transposed prior to input into software.
    h. **PA_DEP_26r_OLI_4majors_min.csv** - Dataset records with at least 4 major cations measured, formatted for input into OLI Studio. Data must be transposed prior to input into software.
    i. **PA_DEP_26r_OLI_charge_balance.csv** - Charge balance calculations for data included in the PA_DEP_26r_OLI.csv file.
    j. **PA_DEP_26r_OLI_charge_balance_1val_min.csv** - Charge balance calculations for data included in the PA_DEP_26r_OLI_1val_min.csv file.
    k. **PA_DEP_26r_OLI_charge_balance_4majors_min.csv** - Charge balance calculations for data included in the PA_DEP_26r_OLI_4majors_min.csv file.
3. **README_PA_DEP_26r_Produced_Water.pdf** (this document) – Metadata including citation information, dataset descriptions, funding sources, and points of contact.

4. **PA_DEP_26r_Field_Dictionary.csv** – Field dictionary with attribute descriptions, reporting units, and data types.
5. **CoDaRT Files**
    a. **PA_DEP_26r_OLI_4majors_min_codart_input.csv -** Input file to CoDaRT for OLI.
    b. **PA_DEP_26r_GWB_4majors_min_codart_input.csv -** Input file to CoDaRT for GWB.
    c. **PA_OLI_processed_reordered_4majors_min_codart_replaced.csv** -The csv output file from CoDaRT with replacement (for OLI). The following features were not included in the model: Sample_ID, Original_Dataset_ID, Original_Latitude, Original_Longitude, TEMPC, NEWTS_Water_Type.
    d. **PA_OLI_processed_reordered_4majors_min_codart_replaced.xlsx -** The excel output file from CoDaRT with replacement (for OLI). The following features were not included in the model: Sample_ID, Original_Dataset_ID, Original_Latitude, Original_Longitude, TEMPC, NEWTS_Water_Type. The excel version of the output file includes highlighting for replaced values.

Data Processing Notes:

The dataset was subjected to an initial cleaning and processing protocol followed by an attribute mapping procedure used to identify common and unique attributes. The attribute mapper was then used as a guide to standardize field names and integrate common attributes using customized Python scripts.

- General Initial Processing Protocol
    - Filtering
        - Data were filtered to include only sample types of interest. For example, data are filtered for only water samples, leaving out gas and solids.
    - Data Reformatting and Cleaning
        - Units and analyte names were cleaned to have consistent formatting and spelling.
        - Commas were removed from field names to accommodate csv format.
        - The pdf scraped dataset was reformatted from “long” to “wide” format so that each row represents a unique sample containing all measurements associated with that sample. Unique samples were identified using a combination of fields including doc_id, file_name, doc_group, company, facility_name, latitude, longitude, sample_id, form_date, collect_date, matrix, and description. Some of these fields have been removed from the published datasets because they contained sensitive identifying information.
    - Handling Detection Limits
        - If a measurement was below the reported detection limit, the value was changed to 0.
    - Unit Processing
        - If alkalinity, hardness, or acidity measurements were reported only as mg/L and we could not confirm whether it was reported as CaCO3, HCO3, etc. it was separated into a new column.
- Attribute Mapping Protocol

    - Column names were matched to other column names with the same meanings and measurements, not units. Units were handled during integration and conversion.
    - Operator-specific fields that would not be interpretable or useful in the context of this dataset were not attribute mapped and are not included in the final dataset.
    - Fields containing sensitive identifying information were excluded from the published datasets.
    - For each mapped attribute, the input format (e.g., double, long, text, date) and units (when applicable) were included when available in the original resource metadata.
    - When non-metallic element concentrations were labeled using the pure element symbol (Br, Cl), they were mapped to the ionic form (bromide, chloride).
- Integration Protocol
    - If a measurement was reported with "<" it was set to 0.
    - If a measurement was reported with ">", the ">" was removed and the associated numeric value was maintained.
    - Units were converted for analytes previously mapped together. For example, an analyte measured in both mg/L and ug/L would be converted into mg/L and combined into one column during integration.
    - If original measured values did not have units reported in the metadata or through the resource, they were considered unfit for conversion and were excluded from the processed dataset.
- Stoichiometric Conversions
    - Conversions were completed on the data to convert analytes into measurements most compatible with geochemical modeling softwares OLI and GWB. For example, uranium was converted into UO_3 for input into OLI and UO_2^{++} for input into GWB. Analytes were multiplied by a ratio of the output species molecular weight over the as measured species molecular weight. To convert mg/L to meq/L for charge balance calculations, species concentration was multiplied by a ratio of the electrical charge over the molecular weight.