This readme file was prepared for the Rural Evidence and Learning for Water (REAL-Water) Program on March 1, 2023 by Ganesh Shinde.

GENERAL INFORMATION

1. Title of dataset:

   Integrated District-level Water Quality and Scarcity Estimates for India.

2. Description of the dataset:

   This dataset combines spatially explicit information on groundwater and surface water stress in India from multiple, publicly available sources into a single district-level shapefile. For groundwater quality, the researchers downloaded data from the Central Ground Water Board (CGWB) for the years 2010, 2012, 2014, 2016, and 2018.

   The researchers obtained groundwater extraction and recharge information from the 2020 National Compilation on Dynamic Ground Water Resources of India. For surface water scarcity, they calculated the extent of changes in surface water using the Global Surface Water Explorer, a remote sensing-derived product of the European Commission's Joint Research Centre (JRC) within the Copernicus Programme framework.

   The dataset also contains district-level aquifer information from the CGWB and district-level rainfall information from Indian Meteorological Department (IMD).

3. Principal Investigator information:

   Name: Veena Srinivasan
   Institution: Ashoka Trust for Research in Ecology and Environment (ATREE)
   Address: Bengaluru - 64, India
   Email: veena.srinivasan@gmail.com

4. Research Associate information:

   Name: Ganesh Shinde
   Institution: ATREE
   Address: Bengaluru - 64, India
   Email: gsnshinde@gmail.com

5. Date of data collection:

   For groundwater quality and exploitation as well as surface water extents, we downloaded data from May 1, to August 15, 2022.

   For precipitation and aquifer properties, we downloaded data from September 1 to October 15, 2022.

6. Geographic location of data collection:

   We collected the data for the entire country of India with the spatial reference coordinates of 97.415293 East, 68.186249 West, 37.078268 North, 6.755953 South.

7. Information about funding sources that supported collection of the data:

This dataset was processed by ATREE from publicly available information for the REAL-Water Program, a centrally funded research mechanism of the United States Agency for International Development, under Cooperative Agreement Number 7200AA21CA00014.

8. Disclaimer:

   This dataset shared was compiled from public data obtained from either Government of India websites or the European Commission's JRC's Global Surface Water Explorer. Neither ATREE nor the REAL-Water Program produced any of the input data; we have only processed them to report district-level statistics. We encourage users of these data to scrutinize the original source inputs and be mindful of the data's limitations (e.g., low numbers of observations per district for some parameters in a given year, or unusual trends that appear to be driven by such non-physical factors as political boundaries).

SHARING/ACCESS INFORMATION

9. Licenses/restrictions placed on the data:

   Public domain.

10. Was data derived from another source? YES

    If yes, list source(s):

    - CGWB, https://cgwb.gov.in/reports.html (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.)
    - Arsenic Contamination (CGWB), https://cgwb.gov.in//ARSENIC.pdf (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.)
    - Groundwater Exploitation (CGWB), http://cgwb.gov.in/documents/2021-08-02-GWRA_India_2020.pdf (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.)
    - Survey of India (SOI), https://onlinemaps.surveyofindia.gov.in/Digital_Product_Show.aspx
    - Earth Engine Data Catalog: JRC Global Surface Water Mapping Layers, v1.4, https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_4_GlobalSurfaceWater#description
    - Principal Aquifer Systems of India: https://www.indiawris.gov.in/wris/#/Aquifer
    - Precipitation Data: https://imdpune.gov.in/cmpg/Griddata/Rainfall_25_NetCDF.html

11. Recommended citation for this dataset:

    REAL-Water. (2023). Integrated District Shapefile. United States Agency for International Development (USAID) Rural Evidence and Learning for Water.

DATA AND FILE OVERVIEW

12. File list:

    Integrated District Shapefile.zip contains a single shapefile (a geospatial vector format developed for geographic information system [GIS] applications by the Environmental Systems Research Institute [ESRI], Redlands, CA USA), with its associated constituent files (.cpg, .dbf, .prj, .sbn, .sbx, .shp, .shx).

13. Description:

    The polygon structure of the dataset are Indian district boundaries, which are freely downloadable from the SOI. The tabular attribute information corresponds to district-level values for groundwater quality, extraction, recharge, and aquifer properties from India's CGWB; surface area extents from the JRC's Global Surface Water Explorer (via Google Earth Engine); and precipitation values from the Indian Meteorological Department. We computed district-level statistics for all groundwater, surface water, and precipitation values using the spatial join and tabular field calculator features in ArcMap 10.4 (ESRI, Redlands, CA).

14. Selected limitations:

    The attribute table of the SOI's district shapefile included district names and spellings inconsistent with the district names and spellings used in the census. To resolve this issue, three additional columns were added in the attribute table to represent consistent district names and spellings. Additionally, we added the census district code for the year 2011.

    Because Indian districts have been split or recombined into new districts over the course of the reporting period for this dataset, some results may not match those of the input datasets for a given year, as it relies on SOI's most recent district boundary dataset.

METHODOLOGICAL INFORMATION

15. Description of methods used for collection/generation of data:

    Water Quality (*other than Arsenic*). We produced this dataset from data at http://cgwb.gov.in/reports.html). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.)

    We downloaded the groundwater monitoring dataset from the CGWB website made available in Water Quality Reports on the Central Ground Water Board's website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) Each record in the CGWB's public dataset represents a sampling point (typically a borewell or hand pump) with columns corresponding to specific water quality parameters (among them pH, turbidity, electrical conductivity, total dissolved solids, iron, manganese, total hardness, alkalinity, individual major ions, chloride, and fluoride). All sampling points in the dataset have associated positional information (latitude/longitudes). This dataset features five water quality parameters (nitrate, fluoride, chloride, electrical conductivity, and total dissolved solids) given the REAL-Water focus on rural drinking water. To manage the number of tabular attributes, the dataset includes groundwater data from alternate years only (2010, 2012, 2014, 2016, and 2018).

Arsenic. The data for arsenic was obtained from the CGWB report "Arsenic Hotspots in Groundwater in India" accessed on July 14, 2022 (source: https://cgwb.gov.in/WQ/ARSENIC.pdf [These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.]). The arsenic data pertains to the year 2018 (Source: https://cgwb.gov.in/documents/2021-08-02-GWRA_India_2020.pdf [These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.]). These data are available with location and block names, but without corresponding latitude/longitude information. To merge this parameter's data with the rest of the groundwater dataset, we relied on the District Census Handbooks for the Census of India, 2011 https://censusindia.gov.in/census.website/data/handbooks) as a reference to place the locations and blocks from the arsenic dataset into the correct districts for each year.

Groundwater exploitation. We downloaded the data for groundwater exploitation from the CGWB report "Dynamic Groundwater Resources of India 2020" accessed on April 4, 2022 (source: http://cgwb.gov.in/documents/2021-08-02-GWRA_India_2020.pdf [These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.]), specifically Annexure III(B), "District-wise Categorization of Blocks/Mandals/Taluks in India (as in 2020)," with total number of assessed blocks in each district that fall under the different CGWB categories of withdrawals relative to recharge ("safe" is <70% of recharge; "semi-critical" is 70-90% of recharge; "critical" is 90-100% of recharge; and "over-exploited" is >100% of recharge) as well as the saline category.

Surface water extents. We derived district water surface water extents using the "JRC Yearly Water Classification History, v1.4" dataset from the Google Earth Engine spatial archive using the waterClass band with a 30m resolution (source: https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_4_YearlyHistory).

Aquifer properties. We downloaded the Principal Aquifer System of India dataset from a CGWB report on "Aquifer Systems of India" accessed on October 9, 2022 (source: https://cgwb.gov.in/AQM/India.pdf [These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.]). There are 14 principal aquifer systems in India and one "unclassified" category in the dataset, which were reclassified into two classes: "unconsolidated" aquifer formations (consisting of alluvium) and "consolidated" formations (consisting of all other hard rock formation categories). We then assigned to each district a value corresponding to the proportion of the district area made up of consolidated aquifer and the proportion made up of unconsolidated aquifer, computed via the Summarize Within operation in ArcGIS Pro 2.8. We downloaded the aquifer shapefile from https://www.indiawris.gov.in/wris/#/Aquifer in 2020.

Precipitation Layer. We obtained a daily time series of rainfall from 2011 to 2021 from the IMD. The primary data source of the data is the 0.25 degree * 0.25 degree gridded dataset from IMD (source: https://imdpune.gov.in/cmpg/Griddata/Rainfall_25_NetCDF.html). We retrieved the raw data using a script in Google Earth engine processed at the district level. The value for the district was assigned the average of all precipitation grid values within it.

16. Methods for processing the data:

For groundwater quality parameters, we grouped sampling points into districts using the Spatial Join feature in ArcMap 10.4. We computed the number of sampling points within each district whose parameter values fall outside of permissible limits set by the Bureau of Indian Standards (2015 amendments to the 2012 Drinking Water Specifications), as well as the total number of observations; the maximum, minimum, average, and median values; and the standard deviation.

For groundwater exploitation, we reported the total number of groundwater assessment units in a district, and the number of assessment units in each CGWB exploitation category ("safe" is <70% of recharge; "semi-critical" is 70-90% of recharge; "critical" is 90-100% of recharge; and "over-exploited" is >100% of recharge), as well as the saline category.

For surface water extent parameters, we calculated a five-year average for the periods 1999 to 2003 and 2016 to 2020, respectively. We then calculated the difference between the two averages to derive the percentage of change in surface water extents between the periods.

For the aquifer layer, the input file is a shapefile with each polygon representing a contiguous formation. There were 15 classifications including the "unclassified" category, as some regions of the country remain unidentified according to government data (as shown in the table below).

aquifer-10	**Reclassification**
Alluvium	Unconsolidated
Basalt	Consolidated
Basement Gneissic Complex	Consolidated
Charnockite	Consolidated
Gneiss	Consolidated
Granite	Consolidated
Intrusive	Consolidated
Khondalites	Consolidated
Laterite	Consolidated
Limestone	Consolidated
Quartzite	Consolidated
Sandstone	Consolidated
Schist	Consolidated
Shale	Consolidated
Unclassified	Consolidated

As a first step to processing the original data, we reclassified the existing categories into two categories: consolidated and unconsolidated based on the type of aquifer (refer to the second column in the table above). We then assigned to each district a value equivalent to the proportion of unconsolidated aquifer and the proportion of consolidated aquifer, based on the above reclassification.

For precipitation data obtained from the IMD, we calculated a multi-year coefficient of variation (CV) of daily rainfall for the monsoon season extending from June till September for the years between 2011 and 2021. We calculated this CV of rainfall at a district level for the entire country.

17. Describe any quality-assurance procedures conducted on the data:

Data quality checks were conducted for the water quality dataset at multiple levels.

- For typographical errors in district names, we cross-checked the dataset manually against the Census of India and CGWB datasets. The final district name was as per the Census of India 2011 dataset.

18. Instrument or software-specific information needed to interpret the data:

Any GIS software application capable of reading ESRI shapefile formats can be used to interpret the data.

CODEBOOK AND VARIABLE DESCRIPTION FOR THE FULL SHAPEFILE

19. Number of variables: 226.
20. Number of cases/rows: 733 + 9 (districts not officially named and marked as "Disputed" in the Survey of India shapefile).
21. Variable List:

Column_Name	Description	Units
State_Name	Name of state	Text
DtCensus_C	District census code according to Census of India, 2011	Number
DtCensus_N	District name according to Census of India, 2011	Text
DtCGWB_N	District name according to CGWB database	Text
DtJJM_N	District name according to JJM database	Text
TotObs_10	Total groundwater quality observations in a district for year 2010	Number
ObsCl_10	Total Chloride observations in a district for year 2010	Number
APLCl_10	No. of Chloride observations above BIS limit for year 2010	Number
MinCl_10	Minimum Chloride value for the district for year 2010	mg/L
MaxCl_10	Maximum Chloride value for the district for year 2010	mg/L
AvgCl_10	Average Chloride value for the district for year 2010	mg/L
MedCl_10	Median Chloride value for the district for year 2010	mg/L
SDCl_10	Standard deviation of Chloride value for the district for year 2010	mg/L
ObsNO3_10	Total Nitrate observations in a district for year 2010	Number
APLNO3_10	No. of district Nitrate observations above BIS limit for year 2010	Number

Column_Name	Description	Units
MinNO3_10	Minimum Nitrate value for the district for year 2010	mg/L
MaxNO3_10	Maximum Nitrate value for the district for year 2010	mg/L
AvgNO3_10	Average Nitrate value for the district for year 2010	mg/L
MedNO3_10	Median Nitrate value for the district for year 2010	mg/L
SDNO3_10	Standard deviation of Nitrate value for the district for year 2010	mg/L
ObsF_10	Total Fluoride Observations in a district for year 2010	Number
APLF_10	No. of Fluoride observations above BIS limit for year 2010	Number
MinF_10	Minimum Fluoride value for the district for year 2010	mg/L
MaxF_10	Maximum Fluoride value for the district for year 2010	mg/L
AvgF_10	Average Fluoride value for the district for year 2010	mg/L
MedF_10	Median Fluoride value for the district for year 2010	mg/L
SDF_10	Standard deviation of Fluoride value for the district for year 2010	mg/L
ObsTDS_10	Total TDS Observations in a district for year 2010	Number
APLTDS_10	No. of TDS observations above BIS limit for year 2010	Number
MinTDS_10	Minimum TDS value for the district for year 2010	mg/L
MaxTDS_10	Maximum TDS value for the district for year 2010	mg/L
AvgTDS_10	Average TDS value for the district for year 2010	mg/L
MedTDS_10	Median TDS value for the district for year 2010	mg/L
SDTDS_10	Standard deviation of TDS value for the district for year 2010	mg/L
ObsEC_10	Total EC Observations in a district for year 2010	Number
APLEC_10	No. of EC observations above BIS limit for year 2010	Number
MinEC_10	Minimum EC value for the district for year 2010	µS/cm
MaxEC_10	Maximum EC value for the district for year 2010	µS/cm
AvgEC_10	Average EC value for the district for year 2010	µS/cm
MedEC_10	Median EC value for the district for year 2010	µS/cm
SDEC_10	Standard deviation of EC value for the district for year 2010	µS/cm
TotObs_12	Total Sample Observations in a district for year 2012	Number
ObsCl_12	Total Chloride Observations in a district for year 2012	Number
APLCl_12	No. of Chloride observations above BIS limit for year 2012	Number
MinCl_12	Minimum Chloride value for the district for year 2012	mg/L
MaxCl_12	Maximum Chloride value for the district for year 2012	mg/L
AvgCl_12	Average Chloride value for the district for year 2012	mg/L
MedCl_12	Median Chloride value for the district for year 2012	mg/L
SDCl_12	Standard deviation of Chloride value for the district for year 2012	mg/L
ObsNO3_12	Total Nitrate Observations in a district for year 2012	Number
APLNO3_12	No. of Nitrate observations above BIS limit for year 2012	Number
MinNO3_12	Minimum Nitrate value for the district for year 2012	mg/L
MaxNO3_12	Maximum Nitrate value for the district for year 2012	mg/L
AvgNO3_12	Average Nitrate value for the district for year 2012	mg/L
MedNO3_12	Median Nitrate value for the district for year 2012	mg/L

Column_Name	Description	Units
SDNO3_12	Standard deviation of Nitrate value for the district for year 2012	mg/L
ObsF_12	Total Fluoride Observations in a district for year 2012	Number
APLF_12	No. of Fluoride observations above BIS limit for year 2012	Number
MinF_12	Minimum Fluoride value for the district for year 2012	mg/L
MaxF_12	Maximum Fluoride value for the district for year 2012	mg/L
AvgF_12	Average Fluoride value for the district for year 2012	mg/L
MedF_12	Median Fluoride value for the district for year 2012	mg/L
SDF_12	Standard deviation of Fluoride value for the district for year 2012	mg/L
ObsTDS_12	Total TDS Observations in a district for year 2012	Number
APLTDS_12	No. of TDS observations above BIS limit for year 2012	Number
MinTDS_12	Minimum TDS value for the district for year 2012	mg/L
MaxTDS_12	Maximum TDS value for the district for year 2012	mg/L
AvgTDS_12	Average TDS value for the district for year 2012	mg/L
MedTDS_12	Median TDS value for the district for year 2012	mg/L
SDTDS_12	Standard deviation of TDS value for the district for year 2012	mg/L
ObsEC_12	Total EC Observations in a district for year 2012	Number
APLEC_12	No. of EC observations above BIS limit for year 2012	Number
MinEC_12	Minimum EC value for the district for year 2012	µS/cm
MaxEC_12	Maximum EC value for the district for year 2012	µS/cm
AvgEC_12	Average EC value for the district for year 2012	µS/cm
MedEC_12	Median EC value for the district for year 2012	µS/cm
SDEC_12	Standard deviation of EC value for the district for year 2012	µS/cm
TotObs_14	Total Sample Observations in a district for year 2014	Number
ObsCl_14	Total Chloride Observations in a district for year 2014	Number
APLCl_14	No. of Chloride observations above BIS limit for year 2014	Number
MinCl_14	Minimum Chloride value for the district for year 2014	mg/L
MaxCl_14	Maximum Chloride value for the district for year 2014	mg/L
AvgCl_14	Average Chloride value for the district for year 2014	mg/L
MedCl_14	Median Chloride value for the district for year 2014	mg/L
SDCl_14	Standard deviation of Chloride value for the district for year 2014	mg/L
ObsNO3_14	Total Nitrate Observations in a district for year 2014	Number
APLNO3_14	No. of Nitrate observations above BIS limit for year 2014	Number
MinNO3_14	Minimum Nitrate value for the district for year 2014	mg/L
MaxNO3_14	Maximum Nitrate value for the district for year 2014	mg/L
AvgNO3_14	Average Nitrate value for the district for year 2014	mg/L
MedNO3_14	Median Nitrate value for the district for year 2014	mg/L
SDNO3_14	Standard deviation of Nitrate value for the district for year 2014	mg/L
ObsF_14	Total Fluoride Observations in a district for year 2014	Number
APLF_14	No. of Fluoride observations above BIS limit for year 2014	Number
MinF_14	Minimum Fluoride value for the district for year 2014	mg/L

Column_Name	Description	Units
MaxF_14	Maximum Fluoride value for the district for year 2014	mg/L
AvgF_14	Average Fluoride value for the district for year 2014	mg/L
MedF_14	Median Fluoride value for the district for year 2014	mg/L
SDF_14	Standard deviation of Fluoride value for the district for year 2014	mg/L
ObsTDS_14	Total TDS Observations in a district for year 2014	Number
APLTDS_14	No. of TDS observations above BIS limit for year 2014	Number
MinTDS_14	Minimum TDS value for the district for year 2014	mg/L
MaxTDS_14	Maximum TDS value for the district for year 2014	mg/L
AvgTDS_14	Average TDS value for the district for year 2014	mg/L
MedTDS_14	Median TDS value for the district for year 2014	mg/L
SDTDS_14	Standard deviation of TDS value for the district for year 2014	mg/L
ObsEC_14	Total EC Observations in a district for year 2014	Number
APLEC_14	No. of EC observations above BIS limit for year 2014	Number
MinEC_14	Minimum EC value for the district for year 2014	µS/cm
MaxEC_14	Maximum EC value for the district for year 2014	µS/cm
AvgEC_14	Average EC value for the district for year 2014	µS/cm
MedEC_14	Median EC value for the district for year 2014	µS/cm
SDEC_14	Standard deviation of EC value for the district for year 2014	µS/cm
TotObs_16	Total Sample Observations in a district for year 2016	Number
ObsCl_16	Total Chloride Observations in a district for year 2016	Number
APLCl_16	No. of Chloride observations above BIS limit for year 2016	Number
MinCl_16	Minimum Chloride value for the district for year 2016	mg/L
MaxCl_16	Maximum Chloride value for the district for year 2016	mg/L
AvgCl_16	Average Chloride value for the district for year 2016	mg/L
MedCl_16	Median Chloride value for the district for year 2016	mg/L
SDCl_16	Standard deviation of Chloride value for the district for year 2016	mg/L
ObsNO3_16	Total Nitrate Observations in a district for year 2016	Number
APLNO3_16	No. of Nitrate observations above BIS limit for year 2016	Number
MinNO3_16	Minimum Nitrate value for the district for year 2016	mg/L
MaxNO3_16	Maximum Nitrate value for the district for year 2016	mg/L
AvgNO3_16	Average Nitrate value for the district for year 2016	mg/L
MedNO3_16	Median Nitrate value for the district for year 2016	mg/L
SDNO3_16	Standard deviation of Nitrate value for the district for year 2016	mg/L
ObsF_16	Total Fluoride Observations in a district for year 2016	Number
APLF_16	No. of Fluoride observations above BIS limit for year 2016	Number
MinF_16	Minimum Fluoride value for the district for year 2016	mg/L
MaxF_16	Maximum Fluoride value for the district for year 2016	mg/L
AvgF_16	Average Fluoride value for the district for year 2016	mg/L
MedF_16	Median Fluoride value for the district for year 2016	mg/L
SDF_16	Standard deviation of Fluoride value for the district for year 2016	mg/L

Column_Name	Description	Units
ObsTDS_16	Total TDS Observations in a district for year 2016	Number
APLTDS_16	TDS observations above BIS limit for year 2016	Number
MinTDS_16	Minimum TDS value for the district for year 2016	mg/L
MaxTDS_16	Maximum TDS value for the district for year 2016	mg/L
AvgTDS_16	Average TDS value for the district for year 2016	mg/L
MedTDS_16	Median TDS value for the district for year 2016	mg/L
SDTDS_16	Standard deviation of TDS value for the district for year 2016	mg/L
ObsEC_16	Total EC Observations in a district for year 2016	Number
APLEC_16	No. of EC observations above BIS limit for year 2016	Number
MinEC_16	Minimum EC value for the district for year 2016	µS/cm
MaxEC_16	Maximum EC value for the district for year 2016	µS/cm
AvgEC_16	Average EC value for the district for year 2016	µS/cm
MedEC_16	Median EC value for the district for year 2016	µS/cm
SDEC_16	Standard deviation of EC value for the district for year 2016	µS/cm
TotObs_18	Total Sample Observations in a district for year 2018	Number
ObsCl_18	Total Chloride Observations in a district for year 2018	Number
APLCl_18	No. of Chloride observations above BIS limit for year 2018	Number
MinCl_18	Minimum Chloride value for the district for year 2018	mg/L
MaxCl_18	Maximum Chloride value for the district for year 2018	mg/L
AvgCl_18	Average Chloride value for the district for year 2018	mg/L
MedCl_18	Median Chloride value for the district for year 2018	mg/L
SDCl_18	Standard deviation of Chloride value for the district for year 2018	mg/L
ObsNO3_18	Total Nitrate Observations in a district for year 2018	Number
APLNO3_18	No. of Nitrate observations above BIS limit for year 2018	Number
MinNO3_18	Minimum Nitrate value for the district for year 2018	mg/L
MaxNO3_18	Maximum Nitrate value for the district for year 2018	mg/L
AvgNO3_18	Average Nitrate value for the district for year 2018	mg/L
MedNO3_18	Median Nitrate value for the district for year 2018	mg/L
SDNO3_18	Standard deviation of Nitrate value for the district for year 2018	mg/L
ObsF_18	Total Fluoride Observations in a district for year 2018	Number
APLF_18	No. of Fluoride observations above BIS limit for year 2018	Number
MinF_18	Minimum Fluoride value for the district for year 2018	mg/L
MaxF_18	Maximum Fluoride value for the district for year 2018	mg/L
AvgF_18	Average Fluoride value for the district for year 2018	mg/L
MedF_18	Median Fluoride value for the district for year 2018	mg/L
SDF_18	Standard deviation of Fluoride value for the district for year 2018	mg/L
ObsTDS_18	Total TDS Observations in a district for year 2018	Number
APLTDS_18	TDS observations above BIS limit for year 2018	Number
MinTDS_18	Minimum TDS value for the district for year 2018	mg/L
MaxTDS_18	Maximum TDS value for the district for year 2018	mg/L

Column_Name	Description	Units
AvgTDS_18	Average TDS value for the district for year 2018	mg/L
MedTDS_18	Median TDS value for the district for year 2018	mg/L
SDTDS_18	Standard deviation of TDS value for the district for year 2018	mg/L
ObsEC_18	Total EC Observations in a district for year 2018	Number
APLEC_18	No. of EC observations above BIS limit for year 2018	Number
MinEC_18	Minimum EC value for the district for year 2018	µS/cm
MaxEC_18	Maximum EC value for the district for year 2018	µS/cm
AvgEC_18	Average EC value for the district for year 2018	µS/cm
MedEC_18	Median EC value for the district for year 2018	µS/cm
SDEC_18	Standard deviation of EC value for the district for year 2018	µS/cm
DtAs_N	Districts with Arsenic detected	Text
TotDt_Bl	Total number of blocks in a district	Number
TotDt_AsBl	Total number of blocks with Arsenic detected	Number
Avg_As	Average Arsenic value for the district	mg/L
GWE_Distri	Districts with groundwater exploitation data	Text
GWE_ObsTot	Total number of groundwater observations in a district	Number
GWE_Obs_Sa	Groundwater observations in a district under “Safe” category	Number
GWE_Obs_Sm	Groundwater observations in a district under “Semi Critical” category	Number
GWE_Obs_Cr	Groundwater observations in a district under “Critical” category	Number
GWE_Obs_Ov	Groundwater observations in a district under “Overexploited” category	Number
GWE_Obs_Sl	Groundwater observations in a district under “Saline” category	Number
JRC_1999	JRC surface water extent in year 1999	Square meters
JRC_2000	JRC surface water extent in year 2000	Square meters
JRC_2001	JRC surface water extent in year 2001	Square meters
JRC_2002	JRC surface water extent in year 2002	Square meters
JRC_2003	JRC surface water extent in year 2003	Square meters
JRC_2016	JRC surface water extent in year 2016	Square meters
JRC_2017	JRC surface water extent in year 2017	Square meters
JRC_2018	JRC surface water extent in year 2018	Square meters
JRC_2019	JRC surface water extent in year 2019	Square meters
JRC_2020	JRC surface water extent in year 2020	Square meters
AvgJRC9903	Average JRC surface water extent for 1999-2003	Square meters
AvgJRC1620	Average JRC surface water extent for 2016-2020	Square meters
AvgJRCDiff	Fraction of change in surface water extent from 1999-2003 to 2016-2020	Number
CV_2011	Coefficient of variation of IMD daily rainfall for 2011 monsoon	Number
CV_2012	Coefficient of variation of IMD daily rainfall for year 2012 monsoon	Number
CV_2013	Coefficient of variation of IMD daily rainfall for year 2013 monsoon	Number

Column_Name	**Description**	**Units**
CV_2014	Coefficient of variation of IMD daily rainfall for year 2014 monsoon	Number
CV_2015	Coefficient of variation of IMD daily rainfall for year 2015 monsoon	Number
CV_2016	Coefficient of variation of IMD daily rainfall for year 2016 monsoon	Number
CV_2017	Coefficient of variation of IMD daily rainfall for year 2017 monsoon	Number
CV_2018	Coefficient of variation of IMD daily rainfall for year 2018 monsoon	Number
CV_2019	Coefficient of variation of IMD daily rainfall for year 2019 monsoon	Number
CV_2020	Coefficient of variation of IMD daily rainfall for year 2020 monsoon	Number
CV_2021	Coefficient of variation of IMD daily rainfall for year 2021 monsoon	Number
ReClass	Reclassified aquifer formations as unconsolidated or consolidated	Text
Consol_ar	Area of consolidated aquifer formations	Square Kilometers
Uncon_ar	Area of unconsolidated aquifer formations	Square Kilometers
Dist_ar	Area of district	Square Kilometers
Consol_p	Percentage of area of the district comprising consolidated aquifer	Number
Uncon_p	Percentage of area of the district comprising unconsolidated aquifer	Number
Cri_Ov_Pct	Percentage of groundwater observations in a district the proportion of assessed units classified as either "Critical" or "Overexploited"	Number

22. Missing data codes:

In the present dataset, wherever there was no observation data, it was marked as "-999" denoting data was not available. The value of 0 denotes that the particular parameter is not present in the district.

23. Other abbreviations used:

Tot = Total

Obs = Observations

APL = Above Permissible Limit

Avg = Average

Med = Median

Min = Minimum

Max = Maximum

SD = Standard Deviation

Cl = Chloride

NO3 = Nitrate

F = Fluoride

TDS = Total Dissolved Solids

EC = Electrical conductivity

As = Arsenic

Dist = District

GWE = Ground Water Extraction

JRC = Joint Research Centre

CV = Coefficient of Variation

ar = area

IMD = Indian Meteorological Department

BIS = Bureau of Indian Standards

μS = microSiemens

24. Detailed description of variables:

a. TotObs_yy

Parameter description:

The total number of observations/assessed units in each district for all the states of India for the specified year that are monitored by the CGWB. The "yy" denotes the last two digits of year of the data collection. The respective years are 2010 represented as "10," 2012 represented as "12," 2014 represented as "14," 2016 represented as "16," and 2018 represented as "18."

Data codes:

Missing data codes - Districts with no sampling points are listed as -999 (data not available).

Data sources:

Data are from the Water Quality Reports on the CGWB's website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) Data were downloaded between April 4 and April 8, 2022.

Data processing and quality assurance:

We conducted quality checks to ensure that district names were consistent with those from the Census of India (2011).

Data limitations:

Users should note that some districts have low numbers of observations, and Bihar, West Bengal, Jharkhand, and a few north-eastern states have data points for only a few of the years. Where there is considerable spatial heterogeneity in land use, soil, and aquifer properties, district averages based on this dataset may not capture real world conditions accurately in localized portions of a district.

b. Chloride, Cl (ObsCl_<yy>, APLCl_<yy>, MinCl_<yy>, MaxCl_<yy>, AvgCl_<yy>, MedCl_<yy>, SDCl_<yy>)

Parameter description:

Chloride concentration in groundwater for the respective years as specified. The “yy” denotes the last two digits of year of the data collection. The respective years are 2010 represented as “10,” 2012 represented as “12,” 2014 represented as “14,” 2016 represented as “16,” and 2018 represented as “18.”

ObsCl_yy: Total number of sampling points in a district where chloride concentrations were assessed by CGWB.

APLCl_yy: Total number of sampling points in a district where chloride was above the Bureau of Indian Standards (BIS) drinking water specification of 250 milligrams per liter (mg/L).

MinCl_yy: The minimum chloride concentration (in mg/L) in a district.

MaxCl_yy: The maximum chloride concentration (in mg/L) in a district.

AvgCl_yy: The average chloride concentration (in mg/L) in a district.

MedCl_yy: The median chloride concentration (in mg/L) in a district.

SDCl_yy: The standard deviation of the chloride concentrations (in mg/L) in a district.

Data codes:

Missing Data Codes - Districts with no observations were marked as “-999” denoting data not available. The value of 0 denotes that the parameter was assessed but not present in the water sample.

Data sources:

Water Quality Reports on the CGWB’s website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) Data were downloaded between April 4 and April 8, 2022.

Data processing and quality assurance:

We compiled chloride observations at the district scale. We calculated the average, median, minimum, maximum, and standard deviation based on the concentration of chloride for all assessed units within a district.

c. Nitrate, NO3 (ObsNO3_yy, APLNO3_yy, MinNO3_yy, MaxNO3_yy, AvgNO3_yy, MedNO3_yy, SDNO3_yy,)

Parameter description:

Nitrate concentrations in groundwater for the respective years as specified. The “yy” denotes the last two digits of year of the data collection. The respective years are 2010 represented as “10,” 2012 represented as “12,” 2014 represented as “14,” 2016 represented as “16,” and 2018 represented as “18.”

ObsNO3_yy: Total number of sampling points in a district where nitrate concentrations were assessed.

APLNO3_yy: Total number of sampling points in a district where nitrate was above the BIS drinking water specification of 45 mg/L.

MinNO3_yy: The minimum nitrate concentration (in mg/L) in a district.

MaxNO3_yy: The maximum nitrate concentration (in mg/L) in a district.

AvgNO3_yy: The average nitrate concentration (in mg/L) in a district.

MedNO3_yy: The median nitrate concentration (in mg/L) in a district.

SDNO3_yy: The standard deviation of the nitrate concentrations (in mg/L) in a district.

Data codes:

Missing Data Codes - Districts with no observations were marked as "-999" denoting data not available. The value of 0 denotes that the particular parameter was assessed but not present in the water sample.

Data sources:

Water Quality Reports on the CGWB's website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) Data were downloaded between April 4 and April 8, 2022.

Data processing and quality assurance:

We compiled nitrate observations at the district scale. We calculated the average, median, minimum, maximum, and standard deviation based on the concentration of nitrate for all assessed units within a district.

d. Fluoride, F (ObsF_yy, APLF_yy, MinF_yy, MaxF_yy, AvgF_yy, MedF_yy, SDF_yy)

Parameter description:

Fluoride concentrations in groundwater for the respective years as specified. The "yy" denotes the last two digits of year of the data collection. The respective years are 2010 represented as "10," 2012 represented as "12," 2014 represented as "14," 2016 represented as "16," and 2018 represented as "18."

ObsF_yy: Total number of sampling points in a district where fluoride concentrations were assessed.

APLF_yy: Total number of sampling points in a district where fluoride was above the permissible limit of 1 mg/L.

MinF_yy: The minimum fluoride concentration (in mg/L) in a district.

MaxF_yy: The maximum fluoride concentration (in mg/L) in a district.

AvgF_yy: The average fluoride concentration (in mg/L) in a district.

MedF_yy: The median fluoride concentration (in mg/L) in a district.

SDF_yy: The standard deviation of the fluoride concentrations (in mg/L) in a district.

Data codes:

Missing Data Codes - No observations were marked as "-999" denoting data not available. The value of 0 denotes that the particular parameter was assessed but not present in the water sample.

Data sources:

Water Quality Reports on the CGWB's website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) Data were downloaded between April 4 and April 8, 2022.

Data processing and quality assurance:

We compiled fluoride observations at the district scale. We calculated the average, median, minimum, maximum, and standard deviation based on the concentration of nitrate for all assessed units within a district.

e. Total Dissolved Solids, TDS (ObsTDS_yy, APLTDS_yy, MinTDS_yy, MaxTDS_yy, AvgTDS_yy, MedTDS_yy, SDTDS_yy)

Parameter description:

Total Dissolved Solids (TDS) concentrations in groundwater for the respective years as specified. The "yy" denotes the last two digits of year of the data collection. The respective years are 2010 represented as "10," 2012 represented as "12," 2014 represented as "14," 2016 represented as "16," and 2018 represented as "18."

ObsTDS_yy: Total number of sampling points in a district where TDS was assessed.

APLTDS_yy: Total number of sampling points in a district where TDS was above the permissible limit of 500 mg/l.

MinTDS_yy: The minimum concentration (in mg/L) of TDS in a district.

MaxTDS_yy: The maximum concentration (in mg/L) of TDS in a district.

AvgTDS_yy: The average concentration (in mg/L) of TDS in a district.

MedTDS_yy: The median concentration (in mg/L) of TDS in a district.

SDTDS_yy: The standard deviation of TDS concentrations (in mg/L) in a district.

Data codes:

Missing Data Codes – We marked districts with no observations as "-999" denoting data not available. The value of 0 denotes that the particular parameter was assessed but not present in the water sample.

Data sources:

Water Quality Reports on the CGWB's website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) We downloaded the data between April 4 and April 8, 2022.

Data processing and quality assurance:

We compiled TDS concentrations at the district scale. We calculated the average, median, minimum, maximum, and standard deviation based on the concentration of TDS for all assessed units within a district.

Data limitations:

Some districts have not reported TDS values. However, TDS concentration can be derived from electrical conductivity concentration.

f. Electrical Conductivity, EC (ObsEC_yy, APLEC_yy, MinEC_yy, MaxEC_yy, AvgEC_yy, MedEC_yy, SDEC_yy)

Parameter description:

Electrical Conductivity (EC) in the groundwater for the respective years as specified. The "yy" denotes the last two digits of year of the data collection. The respective years are 2010 represented as "10," 2012 represented as "12," 2014 represented as "14," 2016 represented as "16," and 2018 represented as "18."

ObsEC_yy: Total number of sampling points in a district where EC concentrations were assessed.

APLEC_yy: Total number of sampling points in a district where EC was above the permissible limit of 1,500 microSiemens per centimeter (µS/cm).

MinEC_yy: The minimum value of EC (in µS/cm) in a district.

MaxEC_yy: The maximum value of EC (in µS/cm) in a district.

AvgEC_yy: The average value of EC (in µS/cm) in a district.

MedEC_yy: The median value of EC (in µS/cm) in a district.

SDEC_yy: The standard deviation of EC values (in µS/cm) in a district.

Data codes:

Missing Data Codes - Districts with no observations were marked as "-999" denoting data not available. The value of 0 denotes that the particular parameter was assessed but not present in the water sample.

Data sources:

Water Quality Reports on the CGWB's website (Ground Water Quality 2010-2018). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) Data were downloaded between April 4 and April 8, 2022.

Data processing and quality assurance:

We compiled the EC values at the district scale. We calculated the average, median, minimum, maximum, and standard deviation based on the value of EC for all assessed units within a district.

g. Arsenic, As (DtAs_N, TotDt_Bl, TotDt_AsBl, Avg_As)

Parameter description:

Arsenic concentrations in groundwater, measured in 2018.

DtAs_N: Total number of districts in each state where arsenic was present (with concentration values above 0.01mg/L).

TotDt_Bl: Total number of blocks in a district, referenced with the Census of India.

TotDt_AsBl: Total number of blocks where arsenic was present in the groundwater.

Avg_As: the average arsenic concentration (in mg/L) in a district.

Data codes:

Missing Data Codes - We marked districts with no observations as "-999" denoting data not available. The value of 0 denotes that the particular parameter was assessed but not present in the water sample.

Data sources:

"Arsenic Hotspots in Groundwater in India" (source: https://cgwb.gov.in/WQ/ARSENIC.pdf). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) We downloaded the data July 14, 2022.

Data processing and quality assurance:

The raw dataset for arsenic gives only the location name and block name along with the concentration of arsenic in the affected districts. We grouped the number of blocks where arsenic was present together at a district level.

Data limitations:

Spatial information (latitude/longitude) were not available for this dataset. Since arsenic is geogenic and arsenic contamination is confined to very specific districts that are also well documented in the literature, the data were judged to be reliable.

h. Groundwater Exploitation, GWE (GWE_Distri, GWE_ObsTot, GWE_Obs_Sa, GWE_Obs_Sm, GWE_Obs_Cr, GWE_Obs_Ov, GWE_Obs_Sl)

Parameter description:

These columns pertain to the level of groundwater exploitation in a district.

GWE_Distri: Districts with available data on level of groundwater exploitation.

GWE_ObsTot: Total number of assessed units in a district for groundwater exploitation.

GWE_Obs_Sa: The number of assessed units classified as "Safe" in a district.

GWE_Obs_Sm: The number of assessed units classified as "Semi-critical" in a district.

GWE_Obs_Cr: The number of assessed units classified as “Critical” in a district.

GWE_Obs_Ov: The number of assessed units classified as “Overexploited” in a district.

GWE_Obs_Sl: The number of assessed units classified as “Saline” in a district.

Cri_Ov_Pct: The proportion of assessed units classified as either “Critical” or “Overexploited” in a district.

Data codes:

Missing Data Codes - We marked districts with no observations as “-999” denoting data not available.

Data sources:

Annexure III(B), “District-wise Categorization of Blocks/Mandals/Taluks in India (as in 2020)” downloaded from a CGWB report on “Dynamic Groundwater Resources of India 2020” (source: http://cgwb.gov.in/documents/2021-08-02-GWRA_India_2020.pdf). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.) We downloaded the data April 4, 2022.

Data processing and quality assurance:

The CGWB subdivided each district into “assessment units,” with each assessment unit falling into one of the following five categories:

- Over-Exploited: Groundwater extraction exceeding the annual replenishable ground water recharge.
- Critical: Groundwater extraction between 90-100% of annual extractable resources available.
- Semi-Critical: Groundwater extraction between 70% and 90% of annual extractable resources available.
- Safe: The stage of groundwater extraction is less than 70%.
- Saline category includes districts where EC is above 5000 µS/cm at 25°C in groundwater (source: https://cgwa-noc.gov.in/landingpage/Guidlines/Saline%20guidelines_approved.pdf).

Annexure III(B) reports the number of assessment units within each district falls into each of these categories. The present dataset reports this tabular data, which is joined with the district shapefile using ArcMap 10.4.

Data limitations:

The numbers mentioned in the above columns under the categories of “safe,” “semi-critical,” “critical,” “over-exploited,” and “saline” are aggregated at a district level and are absolute values. So, for each district, values may be present under all the five categories. Looking at these categories in isolation may not give a true interpretation of the dataset. It is advisable to consider these values in relation to the total number of assessed units while interpreting the data.

i. Surface Water spread area, JRC (JRC_yyyy, AvgJRC9903, AvgJRC1620, AvgJRCDiff)

Parameter description:

Surface water extents using remote sensing data obtained from the JRC of the European Space Agency.

JRC_yyyy: Surface water extent in square meters for different years starting from 1999 till 2020 at a district level.

AvgJRC9903: Average surface water extents for the years 1999 to 2003.

AvgJRC1620: Average surface water extent for the years 2016 to 2020.

AvgJRCDiff: Percentage change in 2016-2020 average surface water extents compared to 1999-2003 average surface water extents.

Data codes:

Missing Data Codes - We marked districts with no observations as "-999" denoting data not available.

Data sources:

"JRC Yearly Water Classification History, v1.4" dataset from Google Earth Engine spatial archive. Source: https://developers.google.com/earth-engine/datasets/catalog/JRC_GSW1_4_YearlyHistory. We downloaded the data between September 4 and September 9, 2022.

Data processing and quality assurance:

For our analysis, we used annual values of district-level surface water area in two sets of five-year data: 1999-2003 (Dataset 1) and 2016-2020 (Dataset 2).

To estimate the change in percentage of area under water bodies between the two periods, an average (five-year average) was calculated for each set. With the resultant average area, we applied the following formula:

= (Dataset 2 - Dataset 1)/Dataset 1

A negative percentage change indicates a decrease in the surface water availability and vice versa. We adopted a multi-year approach to reduce the occurrence of anomalies associated with a single hydrological year.

j. Precipitation Layer (CV_yyyy)

Parameter description:

These columns provide information on the CV of rainfall at a district level received during the four monsoonal months or 122 days (June-September).

CV_2011: Coefficient of variation for 2011 daily monsoon rainfall.

CV_2012: Coefficient of variation for 2012 daily monsoon rainfall.

CV_2013: Coefficient of variation for 2013 daily monsoon rainfall.

CV_2014: Coefficient of variation for 2014 daily monsoon rainfall.

CV_2015: Coefficient of variation for 2015 daily monsoon rainfall.

CV_2016: Coefficient of variation for 2016 daily monsoon rainfall.

CV_2017: Coefficient of variation for 2017 daily monsoon rainfall.

CV_2018: Coefficient of variation for 2018 daily monsoon rainfall.

CV_2019: Coefficient of variation for 2019 daily monsoon rainfall.

CV_2020: Coefficient of variation for 2020 daily monsoon rainfall.

CV_2021: Coefficient of variation for 2021 daily monsoon rainfall.

Data codes:

Missing Data Codes - We marked districts with no observations as "-999" denoting data not available.

Data sources:

The primary data source of the data is the 0.25 x 0.25 degree gridded IMD dataset (source: https://imdpune.gov.in/cmpg/Griddata/Rainfall_25_NetCDF.html).

Data processing and quality assurance:

The raw data provided by the IMD is a spatially interpolated gridded dataset from its network of rain gauge stations spread across the country (for further information: https://web.archive.org/web/20220623040324/https://www.imdpune.gov.in/Clim_Pred_LRF_New/ref_paper_MAUSAM.pdf). Using the *imdlib* python package, we downloaded this gridded dataset and converted it into raster (GeoTiff) format, which was then uploaded to the Google Earth Engine Asset for further analysis. We calculated the CV of daily rainfall at a district level for the entire country. We did this for each year for the monsoon season, extending from June till September from 2011 to 2021.

Data limitation:

For this particular dataset, we have considered only the southeast monsoonal months applicable to India (i.e., June-September). The methodology we adopted for calculating the CV was derived from a similar work by Soman and Kumar, 1990 (source: https://www.tropmet.res.in/awnew/aw-27.pdf).

k. Aquifer Layer (Classi, Consol_ar, Uncon_ar, Dist_ar, Consol_p, Uncon_p)

Parameter description:

These columns describe the area under consolidated and unconsolidated aquifer formations at a district level for the entire country.

- Classi: Reclassified aquifer formations as unconsolidated or consolidated.

- Consol_ar: Area under consolidated aquifer formations (km^2).

- Uncon_ar: Area under unconsolidated aquifer formations (km^2).

- Dist_ar: Total area of the district (km^2).

- Consol_p: Percentage of area of the district under consolidated formations.

- Uncon_p: percentage of area of the district under unconsolidated formations.

Data codes:

Missing Data Codes - We marked districts with no observations as "-999" denoting data not available.

Data sources:

Report on "Aquifer Systems of India" accessed on October 9, 2022 (source: https://cgwb.gov.in/AQM/India.pdf). (These data were available for view and download during the assembly of our integrated dataset, but as of November 2023 they no longer appear on the corresponding website. We are unaware if or when these data may be restored.)

Data processing and quality assurance:

There were 15 classifications, including 14 classes of Principal Aquifer Systems of India and one "unclassified" category. This dataset was reclassified at a district level such that each district falls into one of the two broad categories of "Consolidated" and "Unconsolidated" formations. We then assigned to each district a value corresponding to the proportion of the district area made up of consolidated aquifer and the proportion made up of unconsolidated aquifer, computed via the Summarize Within operation in ArcGIS Pro 2.8.