Chapter 2: Ingestion of Data

The Data Ingestion module of the National Platform for Crop Yield Forecasting (NPCYF) forms the entry point for all externally sourced datasets used in forecasting activities. As a national-scale analytical system, NPCYF depends on diverse data streams—agrometeorological, hydrological, agricultural production, remote sensing, and environmental datasets—to support scientifically rigorous and timely crop yield predictions.

Data ingestion ensures that datasets obtained from official and authenticated sources follow a standardized structure, undergo system-level validation, and are transformed into formats suitable for downstream processing. This chapter outlines the sources, procedures, metadata requirements, and system interfaces involved in transferring raw data into the NPCYF ecosystem.

2.1 Purpose and Scope of the Ingestion Module

The Data Ingestion module is designed to:

- Enable seamless upload of raw datasets obtained externally.
- Ensure structural and semantic validation of uploaded files.
- Capture essential metadata (spatial, temporal, administrative).
- Maintain a traceable lineage from source file → ingested dataset → ETL loaded dataset.
- Prepare data for subsequent phases such as feature engineering, modelling, and forecasting.

The ingestion pipeline handles multiple data categories including APY (Area–Production–Yield), rainfall, temperature, reservoir levels, groundwater, and other agro-environmental indicators.

2.2 Supported Data Categories and Source Systems

NPCYF integrates data from government-backed, certified national and state-level systems. Each dataset category has specific designated sources:

2.2.1 Agricultural Production and Yield (APY) Data

APY datasets constitute the core agricultural statistics that document crop cultivation area (hectares), production volume (tonnes), and productivity (yield per hectare) across crops, districts, states, and time periods. These datasets are essential for training predictive models and validating forecast accuracy.

Unified Portal for Agricultural Statistics (UPAG)

- **Access URL:** https://upag.gov.in
- **Authority:** Ministry of Agriculture and Farmers Welfare, Government of India
- **Data Acquisition Procedure:**
  1. Navigate to the Data Access or Statistical Reports section from the main dashboard
  2. Select the relevant category (Crop Statistics, Horticulture Production, Land Use, etc.)
  3. Choose the specific year(s) and state(s) or all-India aggregation as required
  4. Click "Download Data" to retrieve datasets in Excel (XLSX) or CSV format
  5. Verify file completeness and header structure before ingestion
- **Update Frequency:** Annual releases with quarterly advance estimates for major crops
- **Coverage:** Comprehensive national coverage with district-level granularity

2.2.1.1 Sample APY Data structure allowed in the platform

STATEWISE-APY DATA FORMAT SORTED BY CROP & SEASON (ALL INDIA)

DISTRICTWISE APY FORMAT SORTED BY CROP AND SEASON (ALL INDIA)

Below are the other data sources…

IndiaStat Agricultural Database

- **Access URL:** https://www.indiastat.com/data/agriculture/agricultural-production

- **Authority:** IndiaStat Private Limited (aggregates official government statistics)
- **Data Acquisition Procedure:**
    1. Access the Agriculture section from the main navigation menu
    2. Select the desired dataset category (Production, Yield, Land Use, Crop-wise statistics)
    3. Click "View Data" to preview the dataset structure and available variables
    4. Use "Download Excel/CSV" option (institutional login credentials may be required)
    5. Document the data source, access date, and version for metadata records
- **Update Frequency:** Regular updates aligned with government releases
- **Coverage:** Multi-year time series with state and district breakdowns

Directorate of Economics and Statistics (DES) Agriculture Portal

- **Access URL:** https://data.desagri.gov.in/website/crops-apy-report-web
- **Authority:** Directorate of Economics and Statistics, Ministry of Agriculture
- **Data Acquisition Procedure:**
    1. Navigate to the "Crops - APY Report" section
    2. Select specific crop(s), year(s), and state(s) using the interactive filters
    3. Click "Generate Report" to compile the customized dataset

    4. Download the generated report in Excel or CSV format
    5. Verify data completeness, particularly for recent seasons that may have preliminary estimates
- **Update Frequency:** Seasonal updates with advance, first, and final estimates
- **Coverage:** Crop-specific data with administrative hierarchy (national → state → district)

2.2.2 Rainfall Data

Rainfall is a critical agrometeorological variable that directly influences crop sowing decisions, growth patterns, water availability, and ultimately yield outcomes. The platform ingests historical and real-time rainfall data at multiple temporal and spatial resolutions.

Primary Data Sources:

India Meteorological Department (IMD) - Customized Rainfall Information System (CRIS)

- **Access URL:** https://hydro.imd.gov.in/hydrometweb/(S(…))/landing.aspx#
- **Authority:** India Meteorological Department, Ministry of Earth Sciences
- **Description:** CRIS provides comprehensive rainfall data with customizable temporal (daily, weekly, monthly, seasonal, annual) and spatial (gridded, district, subdivision, state) aggregations
- **Data Acquisition Procedure:**
    1. Access the CRIS portal using institutional credentials
    2. Select the geographical region (district, subdivision, state, or custom boundary)
    3. Choose the temporal resolution and date range
    4. Generate the rainfall report using the query interface
    5. Download the dataset in the preferred format

- **Update Frequency:** Daily updates for current season; historical archives available
- **Coverage:** Pan-India coverage with district-level resolution since 1901

Data.gov.in - Open Government Data Platform

- ***Access URL:*** *https://data.gov.in*
- ***Authority:*** *National Informatics Centre, Government of India*
- ***Data Acquisition Procedure:***
    1. *Enter search terms such as "district wise rainfall" or "rainfall India" in the search interface*
    2. *Filter results to identify datasets from IMD or Ministry of Earth Sciences*
    3. *Review the dataset metadata including temporal coverage, spatial resolution, and update frequency*
    4. *Click the "Download" button to retrieve CSV or Excel files*
    5. *Example relevant dataset: "Monthly Rainfall Data – District-wise" with multi-year coverage*
- ***Update Frequency:*** *Variable depending on specific dataset; typically annual updates*
- ***Coverage:*** *District and subdivision level with monthly/seasonal aggregations*

IndiaStat Meteorological Data Repository

- **Access URL:** https://www.indiastat.com/data/meteorological-data/annual-rainfall
- **Authority:** IndiaStat (aggregates official IMD data)
- **Data Acquisition Procedure:**
  1. Log in using institutional credentials
  2. Navigate to the meteorological data section
  3. Select the required dataset (state-wise annual rainfall, season-wise rainfall, etc.)
  4. Choose specific years and parameters using the filter interface
  5. Click "Download Excel/CSV" to obtain the structured rainfall dataset
- **Update Frequency:** Aligned with IMD official releases
- **Coverage:** State and district level with seasonal breakdowns

2.2.2.1 Sample Rainfall Data structure allowed in the platform(all india district wise)

2.2.3 Temperature Data

Temperature variables, including minimum, maximum, and mean temperature measurements, significantly influence crop phenology, growth rates, pest incidence, and yield potential. The platform integrates both gridded and station-based temperature datasets.

Primary Data Sources:

IMD Gridded Temperature Datasets

- **AccessURL:** https://www.imdpune.gov.in/cmpg/Griddata/Max_1_Bin.html#
- **Authority:** Climate Research and Services, India Meteorological Department
- **Description:** Provides high-resolution gridded temperature data (1°×1° or 0.25°×0.25° resolution) covering minimum temperature, maximum temperature, and mean temperature with monthly and annual aggregations
- **Data Acquisition Procedure:**
    1. Navigate to the gridded data download section
    2. Select the temperature variable (minimum, maximum, or mean)
    3. Choose the temporal resolution (daily, monthly, annual) and spatial coverage
    4. Download the binary or NetCDF format files
    5. Extract state-wise or district-wise aggregations using spatial processing tools
- **Update Frequency:** Monthly updates for recent data; complete historical archives
- **Coverage:** Pan-India coverage from 1951 onwards with uniform spatial resolution

IndiaStat Temperature Time Series

- **AccessURL:** https://www.indiastat.com/data/meteorological-data/temperature
- **Authority:** IndiaStat (official IMD data aggregation)

- **Data Acquisition Procedure:**
    1. Access the temperature data section after login
    2. Select state-wise monthly/annual temperature series
    3. Specify minimum, maximum, or mean temperature variables
    4. Choose the required years and geographical units
    5. Download in Excel or CSV format for direct ingestion
- **Update Frequency:** Aligned with IMD releases
- **Coverage:** State-level aggregations with monthly granularity

2.2.3.1 Sample Temperature Data structure allowed in the platform(all india district wise of both tmin and tmax)

2.2.4 Reservoir Water Level Data

Reservoir storage levels serve as crucial indicators of irrigation water availability, which directly impacts crop production in command areas and influences regional agricultural planning and risk assessment.

Primary Data Sources:

India Water Resources Information System (India-WRIS) / National Water Informatics Centre (NWIC)

- **Access URL:** https://indiawris.gov.in/wris/#/Reservoirs
- **Authority:** Central Water Commission, Ministry of Jal Shakti
- **Data Acquisition Procedure:**
  1. Access the Reservoirs section from the main dashboard
  2. Click on specific state boundaries or individual reservoir markers on the interactive map
  3. View real-time water level, storage volume, capacity percentages, and historical trend graphs
  4. Click "Download Data" (typically located at the bottom of the data table)
  5. Select CSV or Excel format and apply filters for state, date range, or specific reservoirs
  6. Save the downloaded file with appropriate naming conventions including date of acquisition
- **Update Frequency:** Real-time updates (daily or weekly) during crop season
- **Coverage:** Major and medium reservoirs across India with storage capacity data

State Water Resource Department Dashboards

Several states maintain dedicated water resource dashboards that provide more granular, district-level reservoir data with higher update frequencies:

State	Portal URL	Dashboard Section
Maharashtra	https://wrd.maharashtra.gov.in	Reservoir Live Storage Dashboard with real-time updates
Karnataka	https://waterresources.karnataka.gov.in	Dam Level Information with historical comparisons
Tamil Nadu	https://www.tn.gov.in/department/33	Reservoir Level Data section with capacity analysis

Telangana/Andhra Pradesh	http://irrigation.telangana.gov.in	Reservoir Status with project-wise breakdowns

Data Acquisition from State Portals:

1. Navigate to the respective state portal
2. Access the reservoir or dam level information section
3. Select specific reservoirs or download aggregate state data
4. Download available formats (typically Excel, PDF, or CSV)
5. Standardize format to align with NPCYF requirements before ingestion

2.2.5 Groundwater Level Data

Groundwater levels indicate the availability of irrigation water from wells and borewells, which supports crop cultivation in areas without surface irrigation infrastructure. Groundwater data is particularly important for rain-fed agricultural regions.

Primary Data Sources:

India Data Portal (Indian School of Business)

- **AccessURL:** https://indiadataportal.com/p/groundwater/r/mojs-wris_cgwb_wells_level_changes-plot-aaa
- **Authority:** Aggregates data from Central Ground Water Board (CGWB) and state agencies
- **Data Acquisition Procedure:**
    1. Navigate to the groundwater monitoring section
    2. Select the geographical region and temporal parameters
    3. View visualizations and data tables for well level changes
    4. Download the underlying dataset in CSV format
    5. Document the measurement period and well locations for metadata
- **Update Frequency:** Seasonal measurements (pre-monsoon and post-monsoon)
- **Coverage:** Well-level monitoring network across India

State Groundwater Boards and Departments

Individual states maintain dedicated groundwater monitoring networks with more frequent measurements and district-level granularity:

State	Portal URL	Section/Report Name
Maharashtra	https://gsda.maharashtra.gov.in	Groundwater Level Status with district reports
Tamil Nadu	https://wrd.tn.gov.in	Groundwater Monitoring Data with well inventories
Karnataka	https://groundwater.karnataka.gov.in	Groundwater Level Reports with seasonal analysis
Telangana	http://groundwater.telangana.gov.in	Water Level Reports with trend analysis
Gujarat	https://gujwater.gujarat.gov.in	Groundwater Monitoring with quality parameters

Data Acquisition from State Portals:

1. Access the specific state groundwater board portal
2. Navigate to the monitoring data or level status section
3. Select district, block, or monitoring well locations
4. Download seasonal reports or raw measurement data
5. Ensure consistency in measurement units (meters below ground level) before ingestion

Before ingestion, users must ensure that all downloaded files meet system requirements regarding header format, data completeness, and structural uniform.

2.3 Data Ingestion tutorials

2.3.1 APY DATA

→CREATING DATASET:

To create a new dataset, open the Datasets tab.Under Create Dataset, we add a new dataset category name(for example,*title as apy2 ,description-upag_testing,data category-APY)*→ click Create Dataset **.**

→DATA INGESTION

To ingest data, go to the Data Ingestion tab → select the dataset that was previously created → enter the required metadata such as the state name, starting year, and header range*(as shown in the image as an example)* → upload the data file from your local system → and wait for the system to confirm a successful upload.
(NOTE: if multi-header say 2 rows are header then, 1-2 should be the input)

Once complete, the dataset is ready for ETL operations.

→DATA ETL

The ETL process in NPCYF is used to load validated data files into the central database.
To perform ETL, open the **Dashboard** → **ETL** tab → select the required dataset from the dropdown menu (for example, *Dataset - Apy2*) → choose the dataset file (for example, *Crop-Profile_State-Wise*) → and specify the data format (for example, *Advance Estimates*).
Next, select the appropriate APY type and provide an ETL title (for example, *soyabean_kharif_96-24*) → then click **Load into DB** to begin the loading process.

After loading into DB one can view their database.

→DATA REPORTS
Note:data reports are only available for apy datasets

To view summaries of ingested and loaded data, open the Data Reports section → review the dataset status, validation details → and export reports if required.

2.3.2 Rainfall data

→CREATING DATASET:

To create a new dataset, open the Datasets tab.Under Create Dataset, we add a new dataset category name(for example,*title as all india dist wise avg rainfall_2024 ,description-testing,data category-RAINFALL)*→ click Create Dataset

→DATA INGESTION

To ingest data, go to the Data Ingestion tab → select the dataset that was previously created → enter the required metadata such as the state name, starting year, and header range*(as shown in the image as an example)* → upload the data file from your local system → and wait for the system to confirm a successful upload.
(NOTE: if multi-header say 3 rows are header then, 1-3 should be the input)

→DATA ETL

The ETL process in NPCYF is used to load validated data files into the central database.
To perform ETL, open the **Dashboard** → **ETL** tab → select the required dataset from the dropdown menu (for example, *Dataset - all india dist wise rainfall_2024*) → choose the dataset file (for example,rainfall_district_Daily_2000_2023_updated (1)) Next, select the appropriate APY type and provide an ETL title (for example, *rainfall dataset 2000-2023)*→ then click **Load into DB** to begin the loading process.

After loading into DB one can view their database.

2.3.3 Temperature data

→CREATING DATASET:

To create a new dataset, open the Datasets tab.Under Create Dataset, we add a new dataset category name(for example,*title as TEMPERATURE,description-temperature dataset,data category-TEMPERATURE)*→ click Create Dataset

→DATA INGESTION

To ingest data, go to the Data Ingestion tab → select the dataset that was previously created → enter the required metadata such as the state name, starting year, and header range*(as shown in the image as an example)* → upload the data file from your local system → and wait for the system to confirm a successful upload.
(NOTE: if multi-header say 3 rows are header then, 1-3 should be the input)

→DATA ETL

The ETL process in NPCYF is used to load validated data files into the central database.
To perform ETL, open the **Dashboard** → **ETL** tab → select the required dataset from the dropdown menu (for example, *Dataset -Temperature*) → choose the dataset file → Next, select the appropriate APY type and provide an ETL title (for example,min temp daily districtwise 1996-2024*)*→ then click **Load into DB** to begin the loading process.

After loading into DB one can view their database.

2.4 Dataset registration and project association

2.4.1 Dataset Registration

This section defines the dataset category under which external data files will be ingested.
Here, we specify the dataset title, description, and data category such as APY, Rainfall, Temperature, Groundwater, etc.

Registration Procedure:

1. Navigate to the **Datasets** tab from the main navigation menu
2. Click on **Create Dataset** to initiate the registration workflow
3. Enter the **Dataset Title** following naming conventions
4. Provide a detailed **Dataset Description** including source attribution
5. Select the appropriate **Data Category** from the dropdown menu
6. Click **Create Dataset** to save the registration

2.4.2 Project Association

Each dataset must be linked to a project before it can be ingested. Projects are created under Admin Management and allow users to organize datasets by workflow.

Creating project

To create a project, go to Admin Management → Projects. Under Your project→Create Project → enter the project name and a brief description → click Add New Project to save it.

After the project is being created under Associate/Dissociate, a dataset can be linked or removed from any project.This ensures access control and structured dataset management.

2.5 Data ingestion workflow

This tab enables users to upload external data files and define the metadata needed for the platform to interpret them correctly.Here, we select the dataset created earlier and provide required metadata such as state name, starting year, and header range. If multi-row headers are present, the header range must include all header rows (e.g., 1–2 or 1–3 as shown in examples). Users then upload Excel or CSV files obtained from authenticated sources.Upon upload, the system performs validation of headers, formats, and data consistency. If the file passes validation, it becomes available for ETL operations.

Technical Insights

NPCYF follows a modern data lakehouse architecture to manage data efficiently.For further details refer to the appendix.

2.6 Data etl (Extract-Transform-Load)

This tab loads the validated data files into the central NPCYF database.

Here, users select the dataset, choose the ingested file, specify the data type (such as APY Advance Estimates), and provide an ETL title. Once configured, selecting *Load into DB* initiates the ETL process.

After the load completes, a preview table is generated, displaying the processed and structured data. This ensures that variables such as area, production, rainfall, year, and district/state information have been correctly parsed.

After loading into DB one can view their database.

Technical Insights

During the **ETL process**, NPCYF takes raw data from the Data Lake (MinIO) and transforms it into structured, queryable tables in PostgreSQL.For further details refer to the appendix.

2.7 Data Reports

This section provides summary reports specifically for APY datasets.

It displays the status of ingested and ETL-loaded files, verifies the parsed header structure, and presents completeness and validation summaries. Users may review the report to identify issues or export it for documentation.

This enables improved data quality monitoring for APY-related forecasting workflows.

To view summaries of ingested and loaded data, open the Data Reports section → review the dataset status, validation details → and export reports if required.