Automated Quality Control of Observed Weather Station Data

Dr. Ranjini Swaminathan
and Dr. Katharine Hayhoe

Texas Tech University

U.S.-India Partnership for Climate Resilience
Workshop on Development and Applications of Downscaling
Climate Projections

The Challenge of Station Data

Weather station records are often the best long-term data we have – but they do not represent absolute truth.

Errors and uncertainty can be introduced by:

- Lack of standardization - measurements made at different times of day, in different locations, and/or with different equipment
- Instrumentation error – instruments may not be calibrated, or may drift over time, or be affected by a natural or human event
- Human error – recording, selection, and/or reading errors

Types of Errors*

- **Random Errors** are distributed more or less **symmetrically** around zero and do not depend on the measured value. Random errors sometimes result in overestimation and sometimes in underestimation of the actual value. On average, the errors cancel each other out.
- **Systematic Errors** are distributed **asymmetrically** around zero. On average these errors tend to bias the measured value either above or below the actual value. One reason of random errors is a long-term drift of sensors.

*World Meteorological Organization Quality Control Guidelines, 2004

Types of Errors*

- **Large (rough) Errors** are caused by malfunctioning of measurement devices or by mistakes made during data processing; errors are easily detected by checks.
- **Micrometeorological (representativeness) Errors** are the result of small-scale perturbations or weather systems affecting a weather observation. These systems are not completely observable by the observing system due to the temporal or spatial resolution of the observing system. When such a phenomenon occurs during a routine observation, the results may look strange compared to surrounding observations taking place at the same time.

*World Meteorological Organization Quality Control Guidelines, 2004

The Challenge of Station Data

These and other types of errors and uncertainties can lead to:

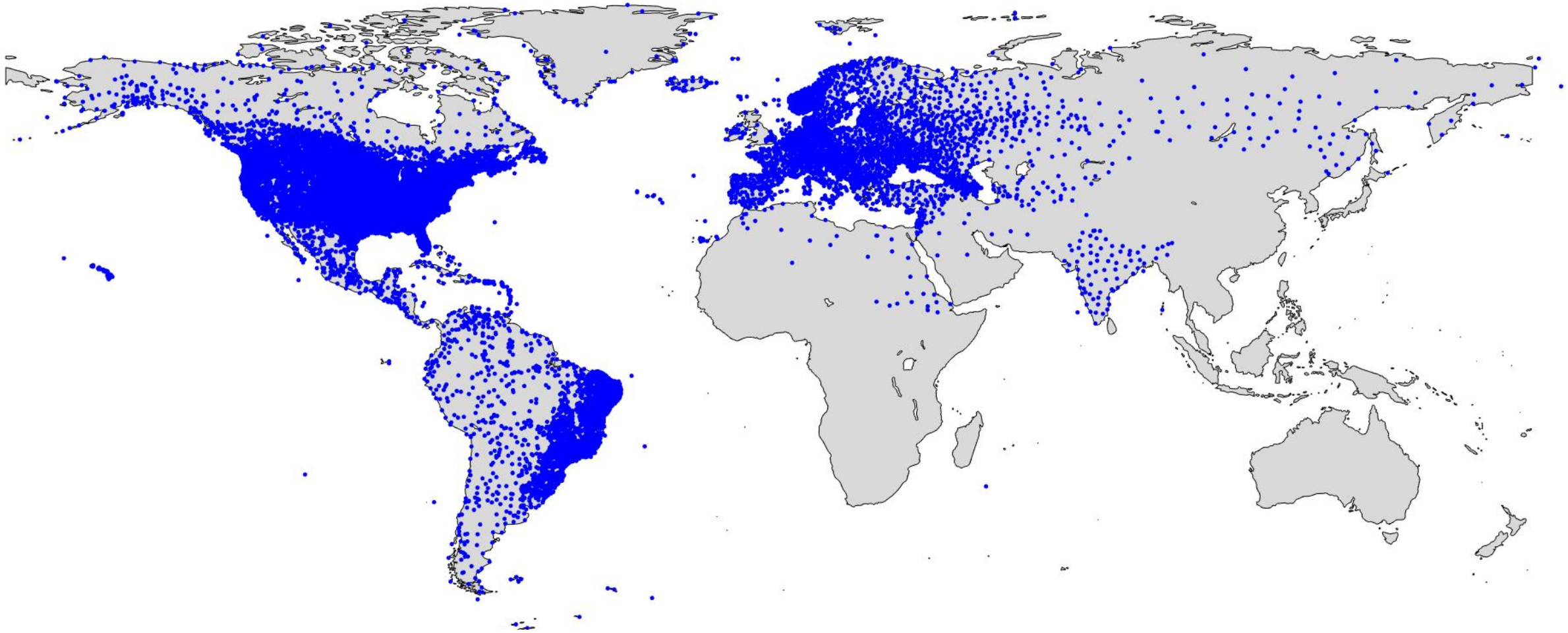
- Missing data
- Unit errors
- Faulty data points
- Records that have been swapped or exchanged
- Other types of biased and/or incorrect records

What station data are we working with?

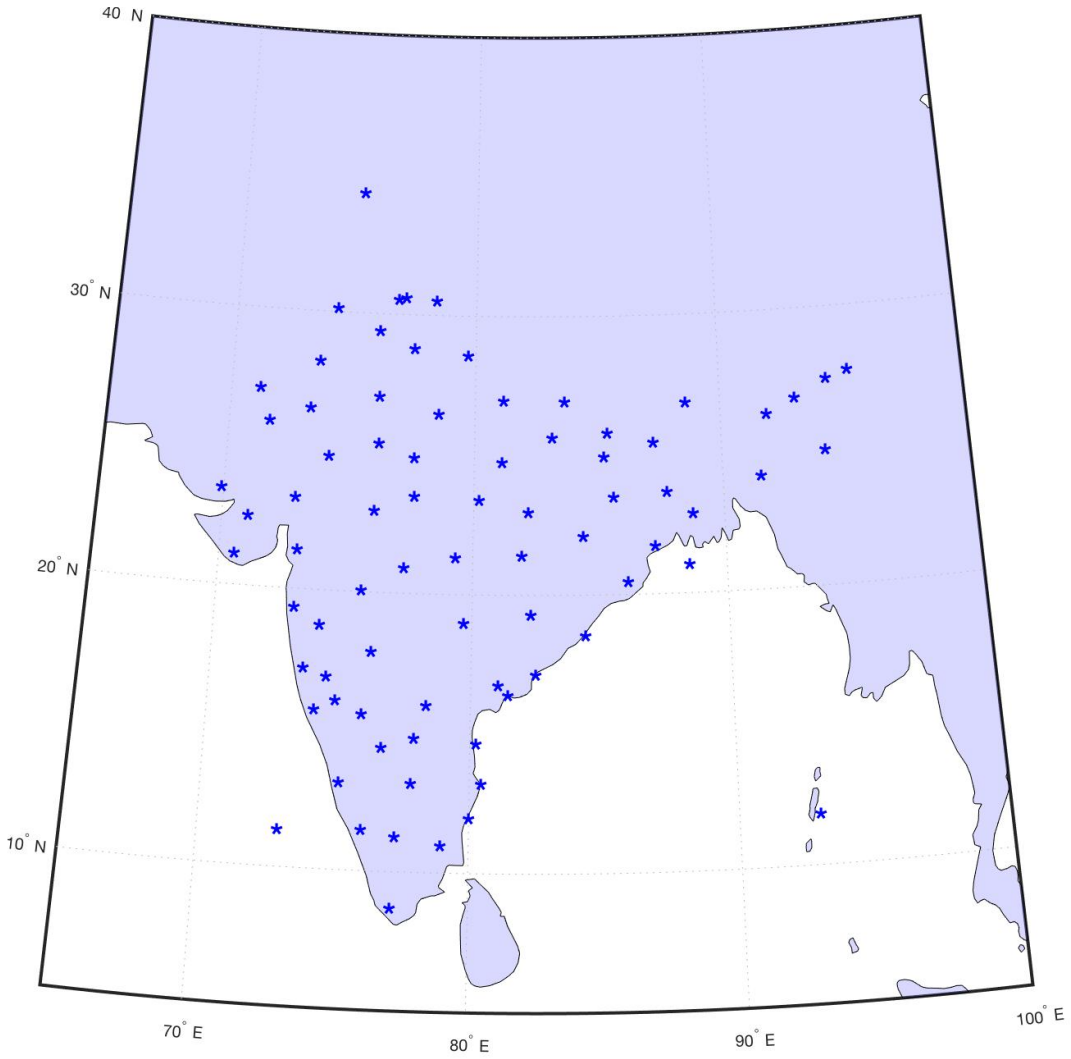
- **Variables** – Daily Temperature (daily maximum, minimum, average) and 24 Hour Cumulative Precipitation

- **Sources** - Global Historical Climatology Network (GHCN), Met Office Integrated Data Archive System (MIDAS), BASINS (Better Assessment Science Integrating point & Non-point Sources), Europe Climate Assessment and Dataset, NOAA and more

Geographic distribution of stations



Long-term weather stations in India

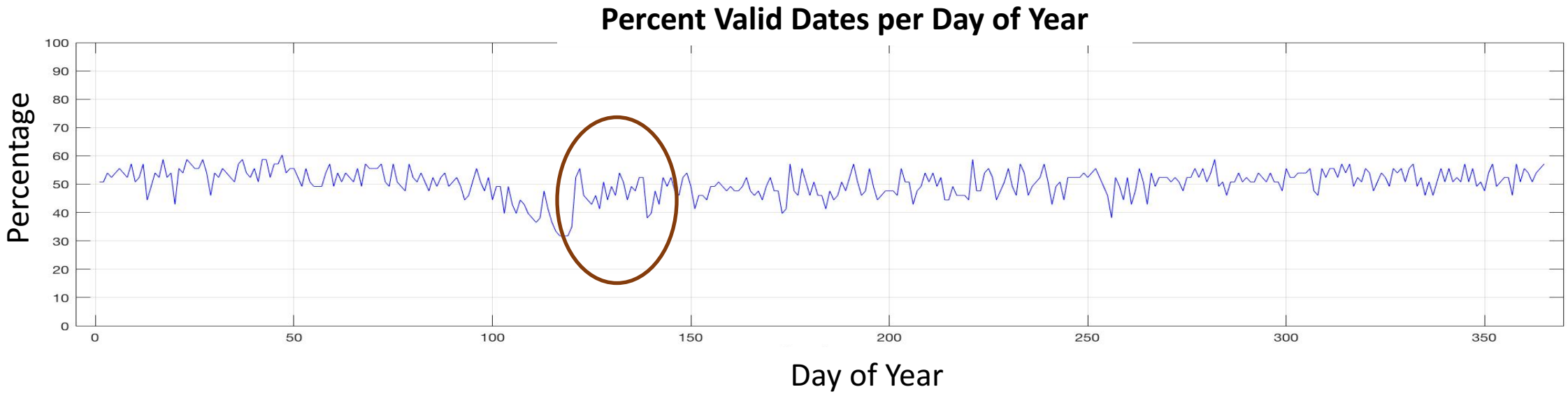


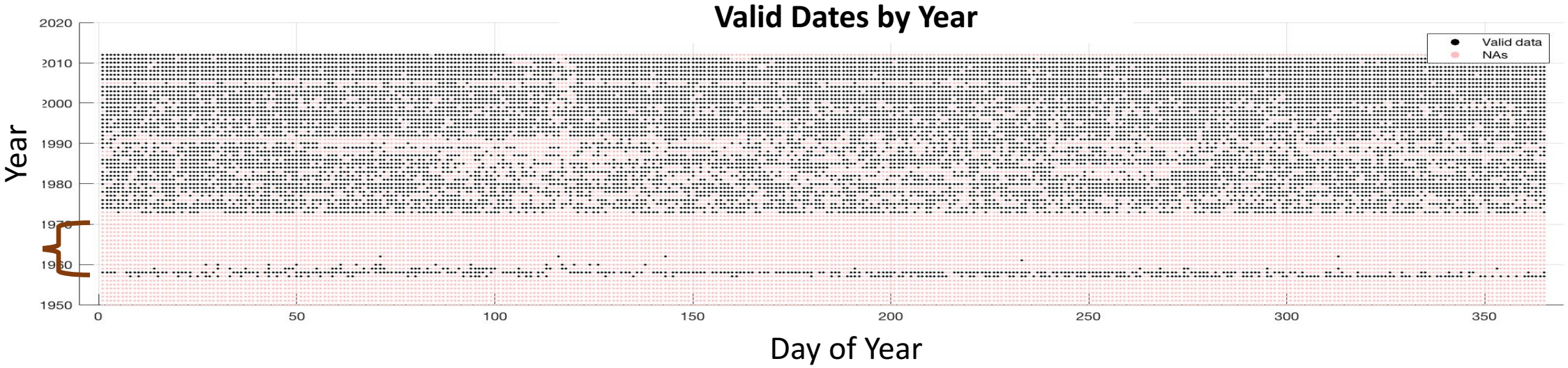
79 weather stations

64 with sufficient long-term data to be used for downscaling

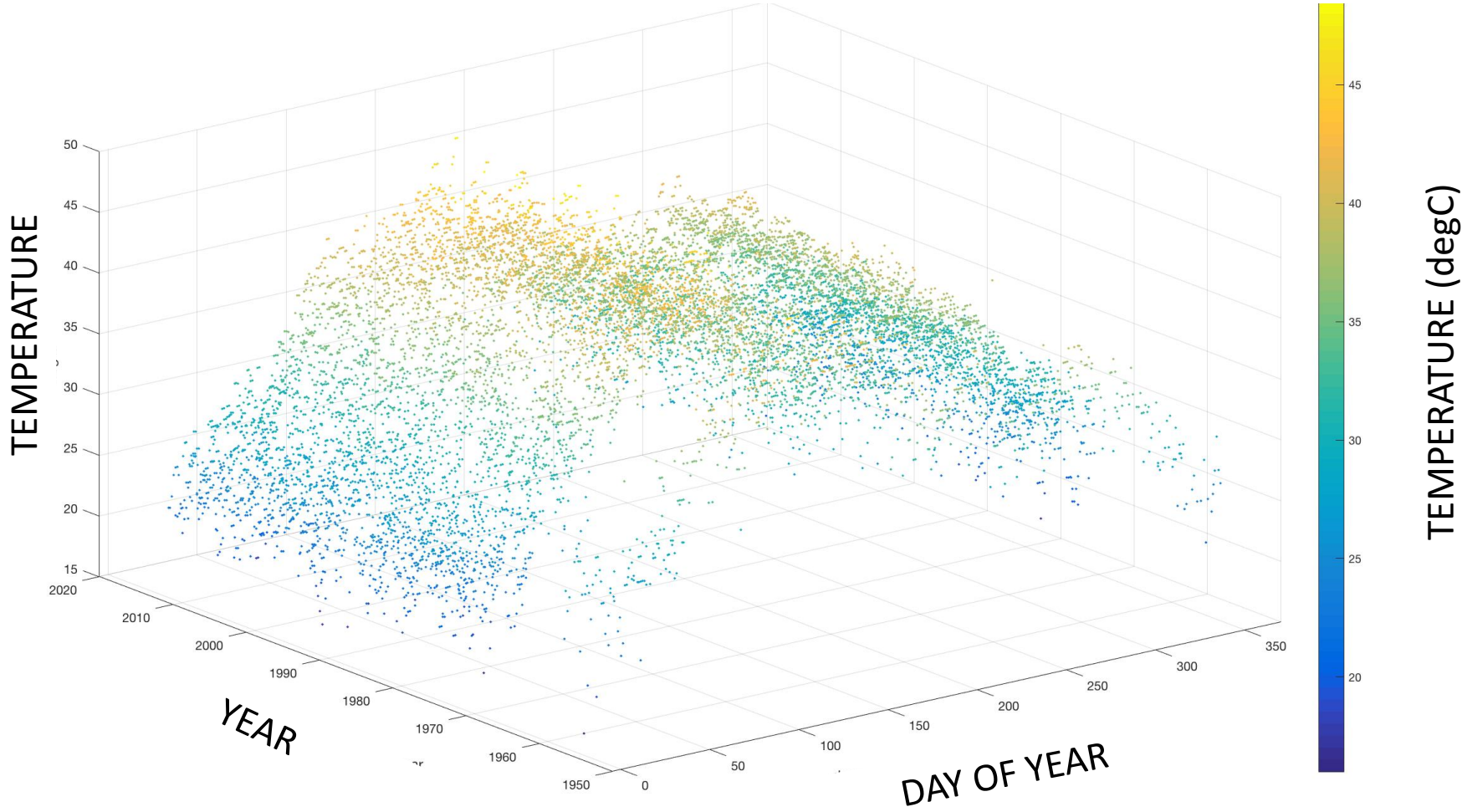
Maximum and minimum temperature, 24 hour cumulative precipitation

Data Availability – Example: Jodhpur





Data Availability – Example: Jodhpur



QA and QC:WMO Definitions

- **Quality Control** – The operational techniques and activities that are used to fulfil requirements for quality.

- **Quality Assurance** -- All the planned and systematic activities implemented within the quality system, and demonstrated as needed, to provide adequate confidence that an entity will fulfil requirements for quality.

*World Meteorological Organization Quality Control Guidelines, 2004

Our QC Method: Basic Checks

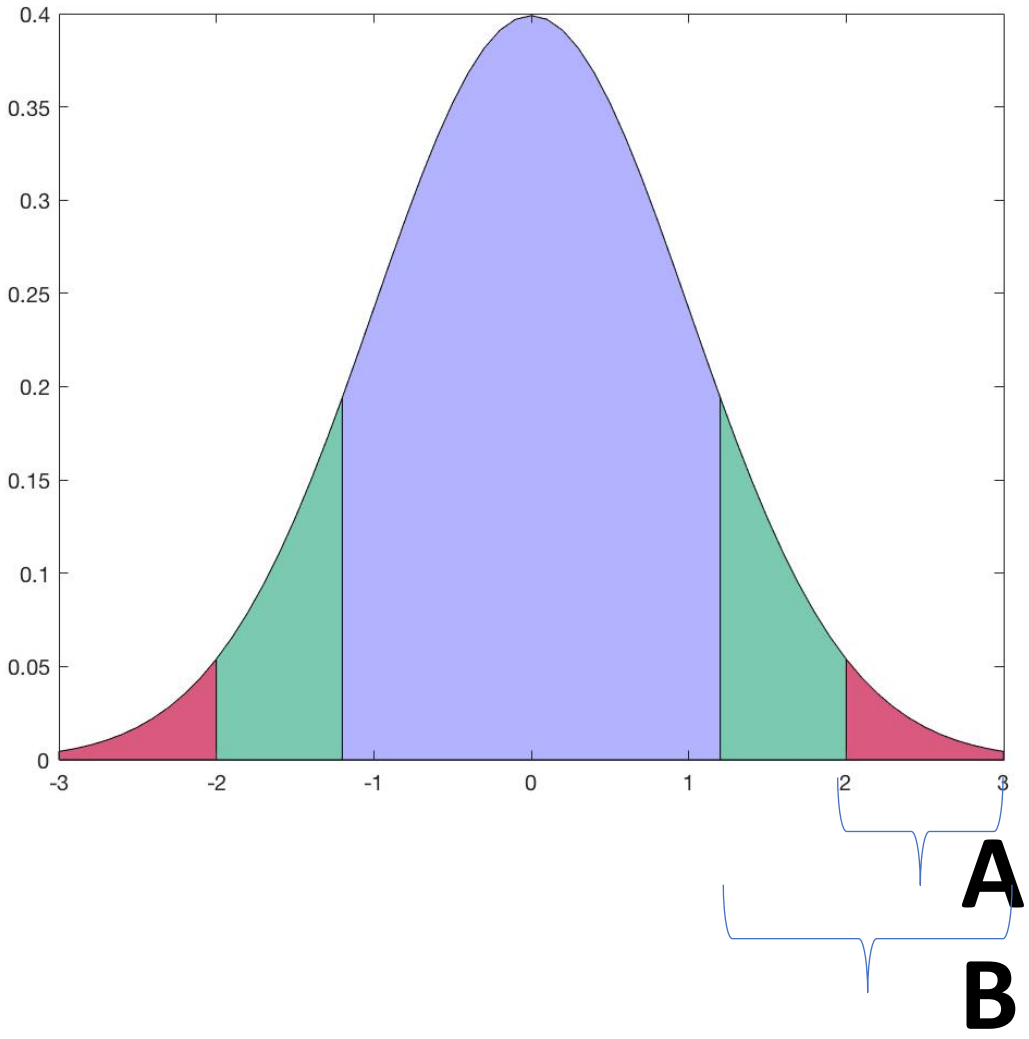
- **Plausible Value Checks :**
  - Temperature and Precipitation – check against maximum and minimum values recorded for country/state/continent.
  - Temperature -- Tmax > Tmin

- **Systemic Error Checks:** Repeated values (5 or more)

Our QC method: Spatio-temporal checks

- Compute nearest neighbors – (neighbor being within a certain geographic distance)
- Temperature data closely follows a normal distribution – we design our metrics for outliers accordingly
- Outliers – outside a certain standard deviation – verified with neighbor highs or lows in a “d” day window period

Spatio-temporal checks



Outliers in region A are investigated

Outliers in region B for “n” nearest neighbor stations and the station itself in a “d” day window period used to validate outliers in region A.

QC Statistics for India Data
TEMPERATURE

- 20 stations lack neighbours for spatio-temporal QC

- Before QC: 59% of TMAX and 63% of TMIN values were missing
- After QC: 60% of TMAX and 64% of TMIN values were missing

- The hottest day on record is 51^{o}C at Phalodi, Rajasthan on 19 May 2016
- The coldest day on record is -33.9^{o}C in Dras on 22 March 1911
- No days exceeded the historical records.

QC Statistics for India Data
TEMPERATURE

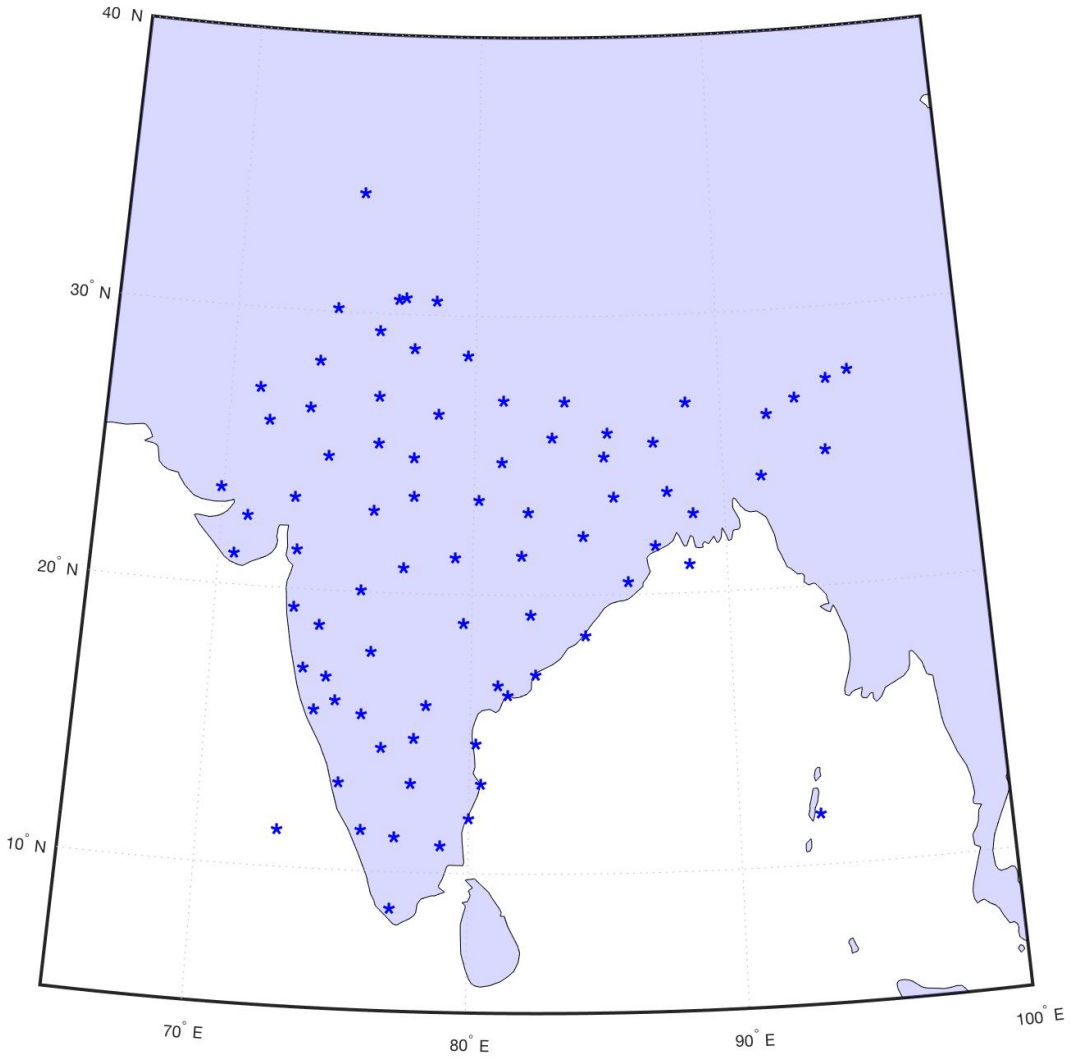
- There were 8 instances of TMIN values exceeding TMAX values at Station IN001050200 (Kakinada). For example:
  - 1987, 11, 11, *23.4 (TMAX)*
  - 1987, 11, 11, *24.3 (TMIN)*

- There were multiple instances of repeated sequences of identical values. The longest repeated sequence was for station IN012070800 (Bombay SantaCruz)
  - From May 6, 1978 to May 20, 1978 – every TMAX value was 33°C

QC Statistics for India Data
PRECIPITATION

- Maximum recorded rainfall – Cherrapunjee (49.05 in, 15th June 1995)
- Number of NA values – same before and after QC (~42%)
- Simple range checks
  - Lower limit : 0 (no negative values)
  - Upper limit: Highest recorded for a given geographic region (state/country)
- No values found to be outside of range

Future climate projections for India



79 weather stations

64 have sufficient long-term data to be used for downscaling

Maximum and minimum temperature, 24 hour cumulative precipitation

In the future, we will be adding:

- **QA index** - quantify quality assurance by categorizing the data as good or erroneous
- **Missing Data Flags** – to determine usability of data
- **Continuous Real Time QC** -- may be able to provide human feedback and to detect and fix source of errors.
- **Topographical Information**

Questions?