# Database Design

## Naming conventions

### General

- All names of DB objects (database, table, column) should use an underscore (_) to separate different properties of that object, and CamelCase should be used to separate words that are part of a phrase.
    - For example, the table temp_cdInputTool_cd uses both. In this case, the variable in question is temperature (temp), the model is the Climate Division Input Tool (cdInputTool), and the type of data is climate division (cd). Note how the underscore (_) separates these 3 descriptive properties. However,CamelCase is used to describe the model, since it's a single property, with multiple words.
- When a table name requires a non-integer (eg. 0.5 month lead), the letters pt will be used. Following this example, the table name for the Monthly Update (lead of 0.5 months) would be temp_manual_0pt5m_01m_stn.
- Numbers will always be zero-padded to two digits (eg. 5 would become 05).

### Forecasts

Wherever possible, tables containing forecast data will follow this naming convention:

```
vvvv_ssss_llu1_wwu2_ttt

where:

  vvvv variable in question (temp, prcp)
  ssss source of the information (manual, auto, tool)
  ll lead time (08, 11...)
  ww averaging window (01, 05, 07...)
  u1 units of lead time (d, m...)
  u2 units of averaging window (d, m...)
  ttt type of data value (cd, stn...)
```

When the source of forecast data is a tool (see the Verification Glossary for a definition of "tool"), ssss will be represented by:

```
xxxx_yyyy_zzz

where:

  xxxx tool that produced the forecast data
  yyyy model that the tool is based on
  zzz  cycle of the model
```

However, when the tool has no cycle (currently only "cdc"):

```
xxxx_yyyy
```

```
where:

  xxxx tool that produced the forecast data
  yyyy model that the tool is based on
```

Again, when the forecast source (ssss) is not a tool, then it is represented by a single string (eg. manual, auto).

- Lead time is always taken to mean the time between the forecast initialization time and the center of the verifying window.
- u1 and u2 are not required to be equal.

## Observations

Wherever possible, tables containing observation data will follow this naming convention:

```
vvvv_wwu_ttt

where:

  vvvv variable in question (temp, prcp)
  ww averaging window (01, 05, 07...)
  u units of averaging window (d, m...)
  ttt type of data value (cd, stn...)
```

## Climatologies

Wherever possible, tables containing climatology data will follow this naming convention:

```
vvvv_wwu_yyyy1_yyyy2_ttt

where:

  vvvv variable in question (temp, prcp)
  ww averaging window (01, 05, 07...)
  u units of averaging window (d, m...)
  yyyy1 Beginning year of climatology
  yyyy2 Ending year of climatology
  ttt type of data value (cd, stn...)
```

## Table Columns

Table columns will follow the follow naming convention:

- The identifier (station ID, WMO#, CD#, etc.) will universally fall under the column name id
- All tables in the observations and climatologies databases will have a column for the valid date, which will be named date_valid.
- All tables in the forecasts database will columns for both the issued date and the valid date, which will be named date_issued and date_valid, respectively.

- Tables in the climatologies database will have a column for the percentiles, as well as the 3 categories. They will follow the convention given in the next bullet point.
- For every column storing the value of the variable in question, the column name will follow the following convention:

  `value_xxx`

  `where xxx is either one of 3 categories:`

  `below`
  `normal`
  `above`

  `or a percentile:`

  `prob_05pcnt`
  `prob_10pcnt`
  `prob_25pcnt`
  `prob_33pcnt`
  `prob_50pcnt`
  `prob_67pcnt`
  `prob_75pcnt`
  `prob_90pcnt`
  `prob_95pcnt`

- Tables in the climatologies database can also contain several parameters defining an approximation of the empirical distribution. They will be named following this convention:

  `param_ppp`

  `where ppp is the parameter in question, such as:`

  `alpha`
  `beta`
  `gamma`

# Database Content

## List of Databases Used

| Name | Purpose |
|------|---------|
| cpc_forecasts | Stores forecast data |
| cpc_observations | Stores observation data |
| cpc_climatologies | Stores climatology data |
| cpc_reference | Stores reference data (station lists, etc.) |
| cpc_vwt | Stores other information needed, such as notes/warnings for the user, etc. |

## Possible Forecast Tables

| Forecast | Lead Times | Averaging Window | Time Units |
|---|---|---|---|
| Extended | D+8, D+11 | 5, 7 | Days |
| Monthly | 0.5, 1 | 1 | Months |
| Seasonal | 2, 3, 4...14 | 3 | Months |

## List of Tables

| Forecast | Example Table Name |
|---|---|
| D+8 | temp_manual_08d_05d_stn |
| D+11 | temp_manual_11d_07d_stn |
| Week 3 | temp_manual_18d_07d_stn |
| Week 4 | temp_manual_25d_07d_stn |
| Monthly (official) | temp_manual_01m_01m_stn |
| Monthly (update) | temp_manual_0pt5m_01m_stn |
| Seasonal (1/2 month lead) | temp_manual_02m_03m_stn |
| Seasonal (1 1/2 month lead) | temp_manual_03m_03m_stn |
| Seasonal (2 1/2 month lead) | temp_manual_04m_03m_stn |
| ... | ... |
| Seasonal (12 1/2 month lead) | temp_manual_14m_03m_stn |
| **Observation** | **Example Table Name** |
| Daily Mean | temp_01d_stn |
| Pentad | temp_05d_stn |
| 7-day Means | temp_07d_stn |
| Monthly Means | temp_01m_stn |
| Seasonal Means | temp_03m_stn |
| **Reference** | **Table Name** |
| Station list | stn |
| Climate Division list | cd |
| Input data settings | dataSettings |
| Date lookup table | dates |

# List of Columns

The following list is a small subset just to give an example of what tables and column names can be expected. For a list of all the tables in a given database, use the Perl script *scripts/db_test_listTables.pl*. Run it with no arguments to print the usage.

**Format of below Database information**
- **Database**
  - *Table*
    - Column

**Databases**
- **climatologies**
  - *temp_07d_1971-2000_cd*
    - id
    - date_valid
    - prob_05pcnt
    - prob_10pcnt
    - prob_25pcnt
    - prob_33pcnt
    - prob_50pcnt
    - prob_67pcnt
    - prob_75pcnt
    - prob_90pcnt
    - prob_95pcnt
    - param_alpha
    - param_beta
    - param_gamma
- **observations**
  - *temp_05d_stn*
    - id
    - date_valid
    - category
- **forecasts**
  - *temp_manual_08d_05d_stn*
    - id
    - date_issued
    - prob_below
    - prob_normal
    - prob_above
- **Reference**
  - *stations*
    - id
    - name
    - state
    - climateRegion
    - climateDivision_102
    - climateDivision_344

- country
- lat
- lon

## Indexes Used

To make querying the database much faster, indexes are used.

### Forecast Tables

| Index Name | Column Indexed |
| --- | --- |
| index_id | id |
| index_date | date_issued |
| index_date_id | (date_issued, id) |

### Observation Tables

| Index Name | Column Indexed |
| --- | --- |
| index_id | id |
| index_date | date_valid |
| index_date_id | (date_valid, id) |

### Climatology Tables

| Index Name | Column Indexed |
| --- | --- |
| index_id | id |
| index_date | date_valid |
| index_date_id | (date_valid, id) |

### Reference Tables

#### Location Lists (stn, cd, grid2deg)

| Index Name | Column Indexed |
| --- | --- |
| index_id | id |

#### Seasonal Signal (climatePhenomena_seasonal)

| Index Name | Column Indexed |
| --- | --- |
| index_date | date_valid |

| index_enso | ENSO |
|---|---|