**数据融合方案**

# 功能描述

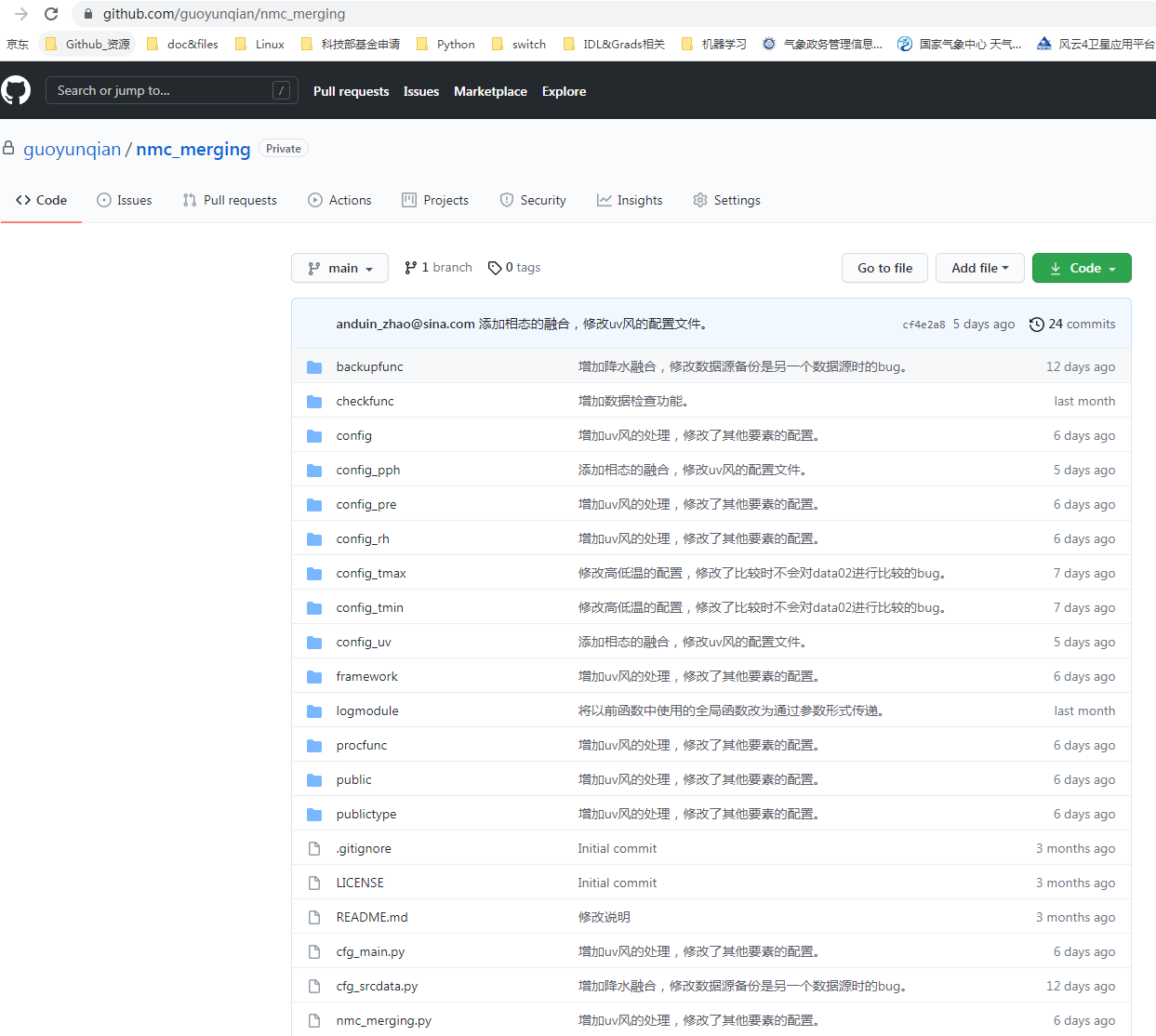
通过融合各个数据源的数据，为app显示提供一份比较准确、比较合理的预报数据，供终端用户查阅使用。

最终的数据源包括但不限于滚动预报、国家指导报、省台预报，当前使用滚动预报、根据省台和国家指导报融合的36小时逐一的指导报和240小时逐三的指导报

代码路径: [\\10.40.17.123\code\nmc\_merging](file:///\\10.40.17.123\code\nmc_merging)

数据路径: [\\10.40.17.123\data](file:///\\10.40.17.123\data)

Github路径: <https://github.com/guoyunqian/nmc_merging>



# 配置信息

配置文件查阅代码根目录中的config目录，主配置文件为cfg\_main.ini，每个数据源各自有一个配置文件，滚动预报的配置文件为cfg\_rp.ini，36小时逐一的融合数据源的配置文件为cfg\_merge\_03601.ini，240小时逐三的融合数据源的配置文件为cfg\_merge\_24003，国家指导报的数据源的配置文件为cfg\_scmoc.ini。

# 处理流程

## 读取配置文件

从主配置文件读出各个数据集的信息、各个处理方法的信息和融合的结果数据的保存信息。根据主配置文件中各个数据集的信息，读取数据集所对应的数据源的信息。

每个数据源的配置包括数据源本身对应的信息、备份信息和检查方法。

备份信息当前包括两种：当前数据源当前起报时的前若干起报时的信息和其他数据源的最近起报时或前若干起报时的信息。备份信息只处理一层，不处理备份数据源的备份信息。在读取备份信息时，生成一个数据源信息列表，数据源本身作为列表中的第一个，然后将备份数据源信息转化为一个个数据源信息保存到列表。一个备份数据源中存在多个起报时间隔的，将每个起报时间隔都生成一个数据源信息，保存到列表。之前讨论的没有数据使用缺失数据的方法，在实际使用中会与数据完整性检查有冲突。

数据检查是确定读取的数据是否有效。当前只有一个通过检查数据中最少存在若干个值，来判断数据是否有效。尚在准备的还有一个检查数据有效范围的方法，通过确定数据是否在某个有效范围，确定数据是否有效，如果无效，还可以根据配置的信息，获得其他数据源的数据进行填充。

## 根据配置文件读取数据

读取每个数据集对应的数据，形成以数据集的名字为key，以从每个数据源读出的数据为value的字典。从数据源读出的数据是以时效为key，以数据为value的字典。

数据集的读取，是根据该数据集前面生成的数据源列表，依次进行。读出的数据要根据配置的检查方法，依次进行检验。建议不通过的数据会被放弃。根据数据源配置的是否完整数据优先，判断何时中断数据源的读取。如果不是完整数据优先，则只要已经读取的所有数据，能组合成一个完整的数据集，读取即告结束，否则一直读取，直到所有数据源都读取过为止。

### 备份方法

#### Forword

Forword方法是指定当前数据源的前若干起报时的数据，作为备份数据。参数是相对于当前所处理的起报时的间隔，可以是多个，以逗号进行分隔。

#### Datasource

Datasource方法指定另外一个数据源作为备份数据。参数是指定数据源的文件名和起报时间隔。起报时间隔是指定数据源的相对于当前处理时间之前的最近的一个起报时，一般是0开始，可以是多个。

### 检查方法

#### diff\_value\_num

diff\_value\_num方法是检查数据中，不同的数值数是否小于指定的值。如果小于指定的值，认为该数据是存在问题的数据，需要放弃。

#### check\_value\_scope（未实现）

check\_value\_scope方法是检查数据中的值是否在指定的有效值范围之内，对于存在值是范围之外的处理，有待商量。

## 对读出的数据进行处理

各个数据集的数据是从数据源中读出的原始数据，并且一些时效的数据可以有多个数据源中的数据，不同时效的数据可以来自不同的数据源，因此这些数据需要进行处理之后，形成一套融合的数据，才能够提供给app使用。

### select

select方法只是从各个读出的数据中，挑选对应时效的数据，对缺失的时效以默认值填充，以组成一个完整的数据源。

### max

max方法是检查一个数据集中的数据，是否不小于另外一个数据集的数据，如果有小于的数据，则替换掉该数据集中对应的数据，最终得到一个包含2个数据集的最大值的数据集。缺测值不参与运算，以有数据的一个为准，如果都是缺测值，则结果集中对应位置为缺测值。

### min

min方法是检查一个数据集中的数据，是否不大于另外一个数据集的数据，如果有大于的数据，则替换掉该数据集中对应的数据，最终得到一个包含2个数据集的最小值的数据集。缺测值不参与运算，以有数据的一个为准，如果都是缺测值，则结果集中对应位置为缺测值。

## 保存数据

经过处理并确定要保存的数据，是以时效为key，数据为value的字典类型的数据，根据保存信息，确定每个时效的数据对应的文件名，然后将数据保存为改名字的文件。文件格式为m4。

# 要素融合

## 温度

### 数据集

温度融合需要的数据集包括滚动预报（data01）、融合插值逐一小时三天的数据（data02）、融合逐三小时十天的数据（data03）和指导报。

所有数据集的备份都是自己前若干个起报时的数据。暂时没有用到指导报。

### 融合

融合的方法是从data01获取1到18的时效，从data02获取19到24的时效，从data03获取27到240的数据。

## 最高温

### 数据集

最高温融合需要的数据集包括融合逐日十天的最高温数据（data01）和温度融合数据（data02）。

所有数据集的备份都是自己前若干个起报时的数据。温度融合数据为前面生成的数据，不需要备份。这也要求温度的融合要放在最高温的融合前面。

### 融合

融合的方法是获取最高温数据和温度融合数据，从结果数据集中挑选出需要的数据，然后获得2个数据集的最大值。

## 最低温

### 数据集

最低温融合需要的数据集包括融合逐日十天的最低温数据（data01）和温度融合数据（data02）。

所有数据集的备份都是自己前若干个起报时的数据。温度融合数据为前面生成的数据，不需要备份。这也要求温度的融合要放在最低温的融合前面。

### 融合

融合的方法是获取最低温数据和温度融合数据，从结果数据集中挑选出需要的数据，然后获得2个数据集的最小值。

## 降水

### 数据集

降水融合需要的数据集包括指导报逐小时一天的数据（data01）、指导报逐三小时十天的数据（data02）、指导报逐小时三天的数据。

所有数据集的备份都是自己前若干个起报时的数据，data01的备份还有指导报逐三小时三天的数据。

### 融合

融合的方法是从data01获取1到24的时效，从data02获取27到240的数据。

## 相对湿度

### 数据集

相对湿度融合需要的数据集包括滚动预报逐小时一天的数据（data01）、融合逐三小时十天的数据（data02）、指导报逐小时三天的数据。

所有数据集的备份都是自己前若干个起报时的数据，data01的备份还有指导报逐三小时三天的数据。

### 融合

融合的方法是从data01获取1到24的时效，从data02获取27到240的数据。

## UV风

### 数据集

UV风的数据集包括融合的逐三小时十天的uv风数据和指导报逐三小时十天的uv风数据。以融合数据的u为data01，v为data02。

所有数据集的备份都是自己前若干个起报时的数据。融合的数据还以指导报对应的数据作为备份。

### 融合

融合的方法是通过数据源获得对应的u和v的数据，通过select\_uv方法处理成对应格式的xarray.DataArray格式的数据。和meteva中处理uv风的结构一致。

## 相态

### 数据集

相态融合需要的数据集包括融合逐三小时十天的相态数据（data01）和指导报逐三小时十天的相态数据。

所有数据集的备份都是自己前若干个起报时的数据。融合的数据集还使用指导报的数据做备份。

### 融合

融合的方法是获取最低温数据和温度融合数据，从结果数据集中挑选出需要的数据，然后获得2个数据集的最小值。