Comparison of efficiency of decision trees and neural networks for postprocessing of thunderstorms

A.Kolker, A. Gochakov, M.Zdereva, N.Hluchina, V. Tokarev

Siberian Regional Meteorological Research Institute, Roshydromet

Novosibirsk,Russia

SibNigmi (RHM) team

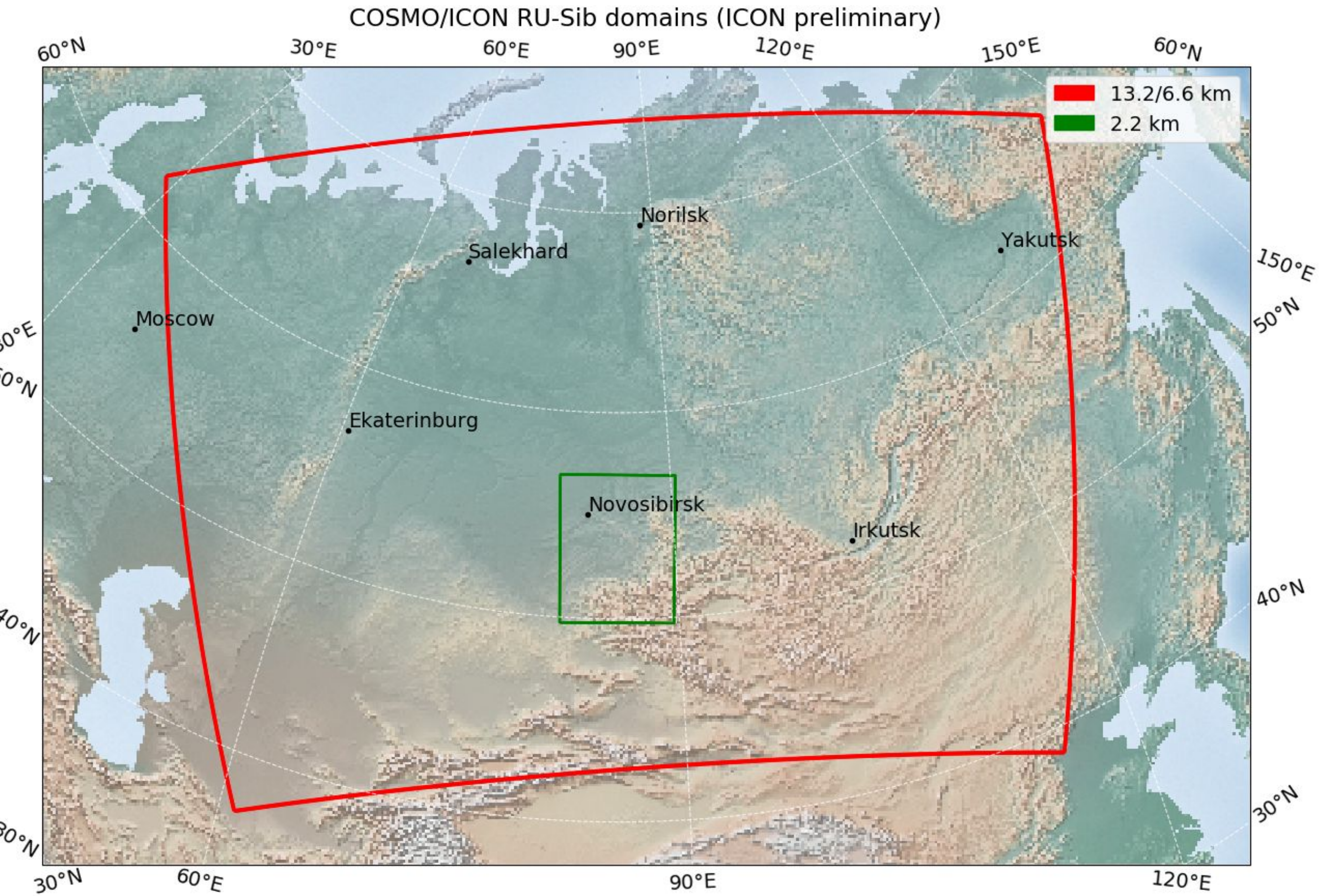
The new team from Siberian Research Meteorological Institute (Novosibirsk, Russia, Roshydromet) has joined to WG4 on July 2021 .

The team has involved to process of investigation on applying ML technologies for rare phenomena (e.g. thunderstorms) forecasting.

The current topic of investigation is learning of effectivity applying Decision Tree and Neural Network classifiers for forecasting thunderstorms (for non-convective Cosmo Model mode) and comparing with direct output of model with convective mode being operated (LPI COSMO and ICON output has analyzed).

The experience could be extended to other type of rare phenomenons.

Data and methods



COSMO v 5.03 (13.2 km), v 5.09 (6.6, 2.2 km)

ICON v 2.6.2.2

Direct variable: LPI (>= 2 J/kg) from 2.2 domains

NN (Sequential, 2 hidden layers) and ML(Decision Tree) 42 variables (direct model output and calculations) from 13.2 COSMO domain

Why we did started working on the topic

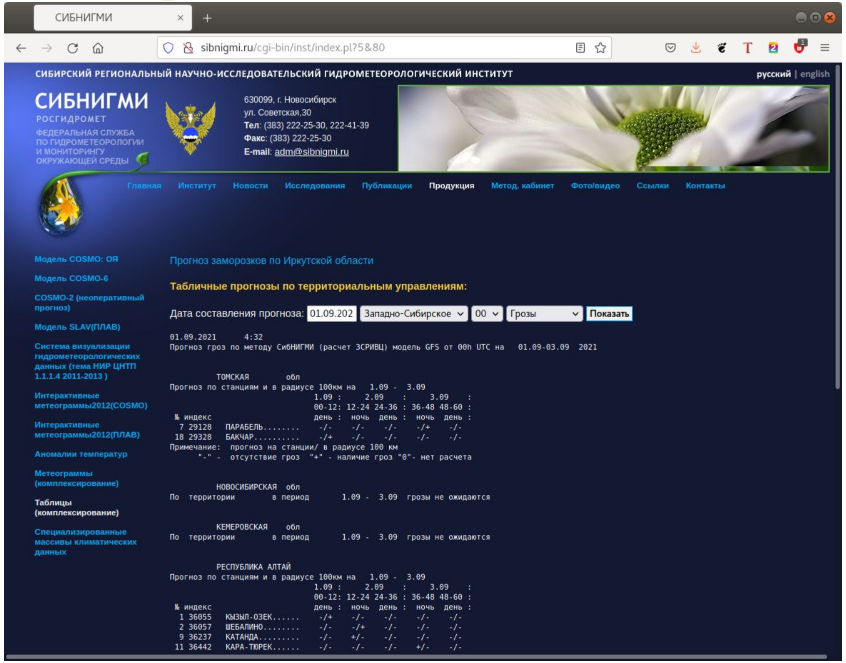
There are Decision Tree based thunderstorm forecast products being operated in our center.

How it constructed:

- 42 variables: direct output of model (COSMO 13.2) extended with:
    - Gradients of dew points,pressure,
    - Laplassian,
    - dew-point deficit value (by lauers)
    - wet-bulb temperature (moisture thermometer),
    - pseudopotential temperature and differences of variables.

Forecast is available on our web server (sorry for inconvenience: thunderstorms are available in Russian page version only)

http://http://sibnigmi.ru/cgi-bin/inst/index.pl?5&80


СИБНИГМИ
sibnigmi.ru/cgi-bin/inst/index.pl?5&80
русский | english
Главная
Институт
Новости
Исследования
Публикации
Продукция
Метод. кабинет
Фото/видео
Ссылки
Контакты
Модель COSMO: ОЯ
Модель COSMO-6
COSMO-2 (неоперативный прогноз)
Модель SLAV(ПЛАВ)
Система визуализации гидрометеорологических данных (тема НИР ЦНТП 1.1.1.4 2011-2013)
Интерактивные метеограммы2012(COSMO)
Интерактивные метеограммы2012(ПЛАВ)
Аномалии температур
Метеограммы (комплексирование)
Таблицы (комплексирование)
Специализированные массивы климатических данных
Показать

СИБИРСКИЙ РЕГИОНАЛЬНЫЙ НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ГИДРОМЕТЕОРОЛОГИЧЕСКИЙ ИНСТИТУТ

СИБНИГМИ
РОСГИДРОМЕТ
ФЕДЕРАЛЬНАЯ СЛУЖБА ПО ГИДРОМЕТЕОРОЛОГИИ И МОНИТОРИНГУ ОКРУЖАЮЩЕЙ СРЕДЫ

630099, г. Новосибирск
ул. Советская,30
Тел: (383) 222-25-30, 222-41-39
Факс: (383) 222-25-30
E-mail: adm@sibnigmi.ru

Прогноз заморозков по Иркутской области

Табличные прогнозы по территориальным управлениям:

Дата составления прогноза: 01.09.202 Западно-Сибирское 00 Грозы

```
01.09.2021      4:32
Прогноз гроз по методу СибНИГМИ (расчет ЗСРИВЦ) модель GFS от 00h UTC на   01.09-03.09  2021


             ТОМСКАЯ        обл
Прогноз по станциям и в радиусе 100км на   1.09 -  3.09
                                       1.09 :     2.09    :     3.09    :
                                       00-12: 12-24 24-36 : 36-48 48-60 :
 № индекс                              день :  ночь  день :  ночь  день :
  7 29128   ПАРАБЕЛЬ........     -/-    -/-    -/-    -/+    -/-
 18 29328   БАКЧАР..........     -/+    -/-    -/-    -/-    -/-
Примечание:  прогноз на станции/ в радиусе 100 км
        "-" -  отсутствие гроз  "+" - наличие гроз "0"- нет расчета


             НОВОСИБИРСКАЯ  обл
По  территории        в период       1.09 -  3.09  грозы не ожидаются


             КЕМЕРОВСКАЯ    обл
По  территории        в период       1.09 -  3.09  грозы не ожидаются


             РЕСПУБЛИКА АЛТАЙ
Прогноз по станциям и в радиусе 100км на   1.09 -  3.09
                                       1.09 :     2.09    :     3.09    :
                                       00-12: 12-24 24-36 : 36-48 48-60 :
 № индекс                              день :  ночь  день :  ночь  день :
  1 36055   КЫЗЫЛ-ОЗЕК......     -/+    -/-    -/-    -/-    -/-
  2 36057   ШЕБАЛИНО........     -/-    -/+    -/-    -/-    -/-
  9 36237   КАТАНДА.........     -/-    +/-    -/-    -/-    -/-
 11 36442   КАРА-ТЮРЕК......     -/-    -/-    -/-    +/-    -/-
```

The targets and value

- To investigate frames of usability ML technologies for postprocessing and rare phenomena forecasting.
- To compare effectivity of various approaches of ML.
- To develop recommendation for training dataset building (e.g. balancing, number of cases and events)

The current status: under processing.

Some details of existing method

1. Only SYNOP observations were used. Forecasts is available only for Synop station location both fixed radius.
2. The system based on 5-years archive of COSMO model. Upgrading of COSMO model leed to necessity of tree rebuilding.
3. The tree was built both mathematics both magic (some changes were manually made to the state of the tree and points used.
4. Every forecast point require own tree. Using someone else's tree is prohibited

Research status and targets

- To build scalable method for building weights for NN and/or Decision Trees
- To investigate efficiency of using various ML approaches for rare phenomena forecasting (classification task).
- To estimate frames usability of various approaches.
- To understand value of direct-convective mode variables.

Source data for training and verification:

- SYNOP, Airep Special , {*Satellites, Radar, Lightening finder*}

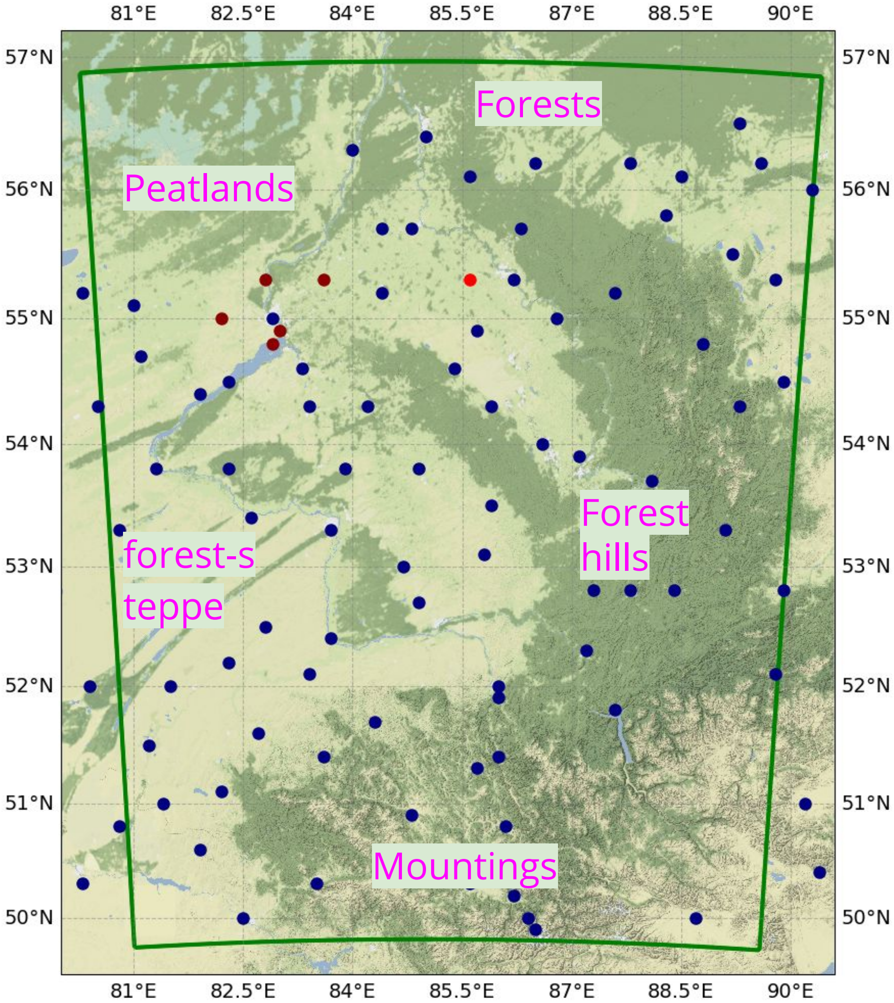
NN structure

NN: Tensorflow Keras Sequental model

- 42 - 52 params input (same for Decision Tree) with normalisation.
- 2 hidden layers with 128 wires and “Relu” activation.
- sigmoid function for final layer.

To be done:

- NN require correct normalisation for operation (default layer was used)



Types of geographical conditions:

- forests,
- peatlands,
- forest-hills,
- forest-steppe,
- mountings

1+1 month experiment (2km-domain area) 0.3 balanced

Trained on: July 2021 set:

yes:1667

no:4228

Tested on: August 2021 set:

yes: 1423

no: 21084

Tree (10 depth)

```
[[  548   875]
 [ 4072 17012]]
```

Total:22507

Hits:18709 (83%)

Hits yes:548 (40%)

hits no:17012 (81%)

False alarms:4072 (286%)

Misses:875 (61%)

q:0.961 p:0.819 H:0.953

NN (150 epoch)

```
[[  576   847]
 [ 2951 18133]]
```

Total:22507

Hits:19325 (83%)

Hits yes:576 (40%)

hits no:18133 (86%)

False alarms:2951 (207%)

Misses:847 (59%)

q:0.962 p:0.869 H:0.957

1+1 month experiment (2km-sized domain) no balancing

Trained on: July 2021 set:

yes:1667

no:21143

Tested on: August 2021 set:

yes: 1423

no: 21084

Tree (10 depth)

```
[[  249  1174]
 [ 1591 19493]]
```

Total:22507

Hits:18709 (88%)

Hits yes:249 (17%)

hits no:19493 (92%)

False alarms:1591 (111%)

Misses:1174 (83%)

q:0.961 p:0.819 H:0.953

NN (150 epoch)

```
[[   14  1409]
 [  155 20929]]
```

Total:22507

Hits:19325 (83%)

Hits yes:14 (1%)

hits no:19493 (99%)

False alarms:155 (11%)

Misses:1409 (99%)

q:0.937 p:0.993 H:0.937

5 year 6+1 stations cluster experiment

Trained on: Summertime +18,+30 2014-2021, 29626,29631,29632,29635,29638,29641 - magenta points on fig, page 9.

yes: 859, no: 6615

Tested on: summertime +18, +30 2014-2021

29641 : red point on fig, page 9:

no: 1327, yes: 167

1. **5 year normalisation**
2. **Limited area with similar geography**

Tree (10 depth)

```
[[  72   95]
 [  72 1255]]
```

Total:1494

Hits:1327 89%

Hits yes:**43%**

Hits no:95%

False alarms:72 **43%**

Misses:95 66%

q:0.936 p:0.952 H:0.933

NN (150 epoch)

```
[[  73   94]
 [  69 1258]]
```

Total:1494

Hits:1331 89%

Hits yes:**44%**

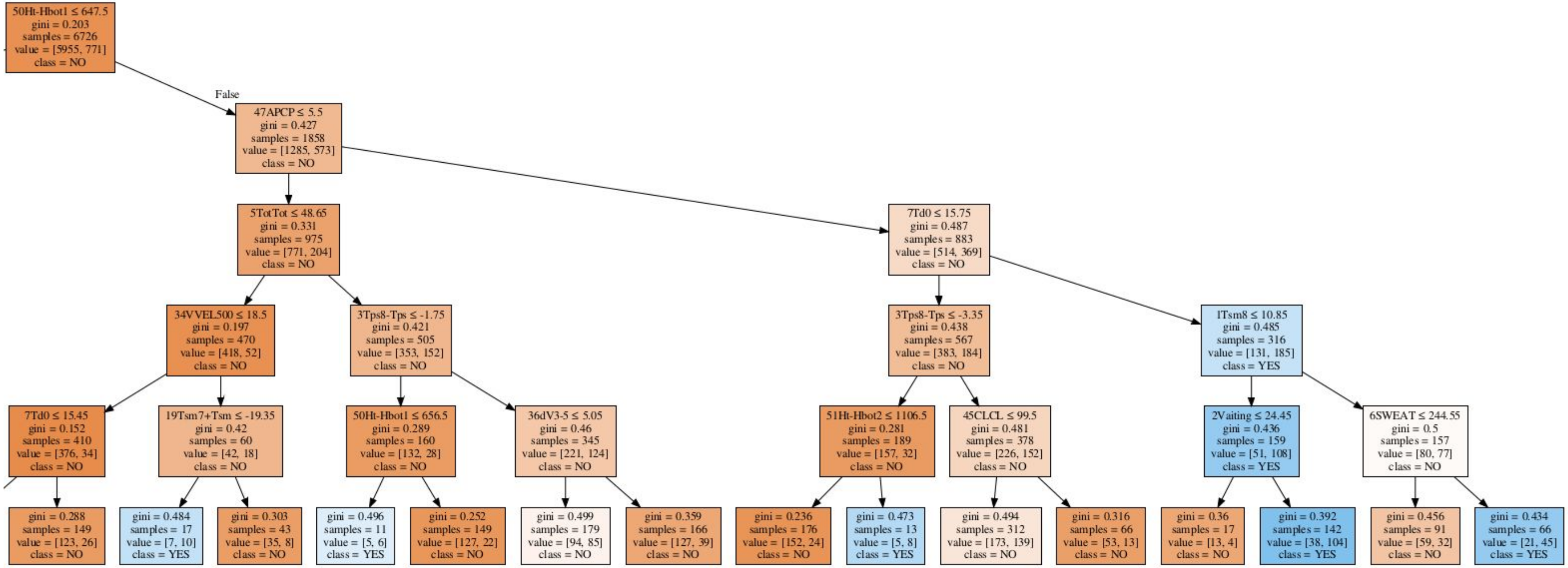
Hits no:95%

False alarms:69: **41%**

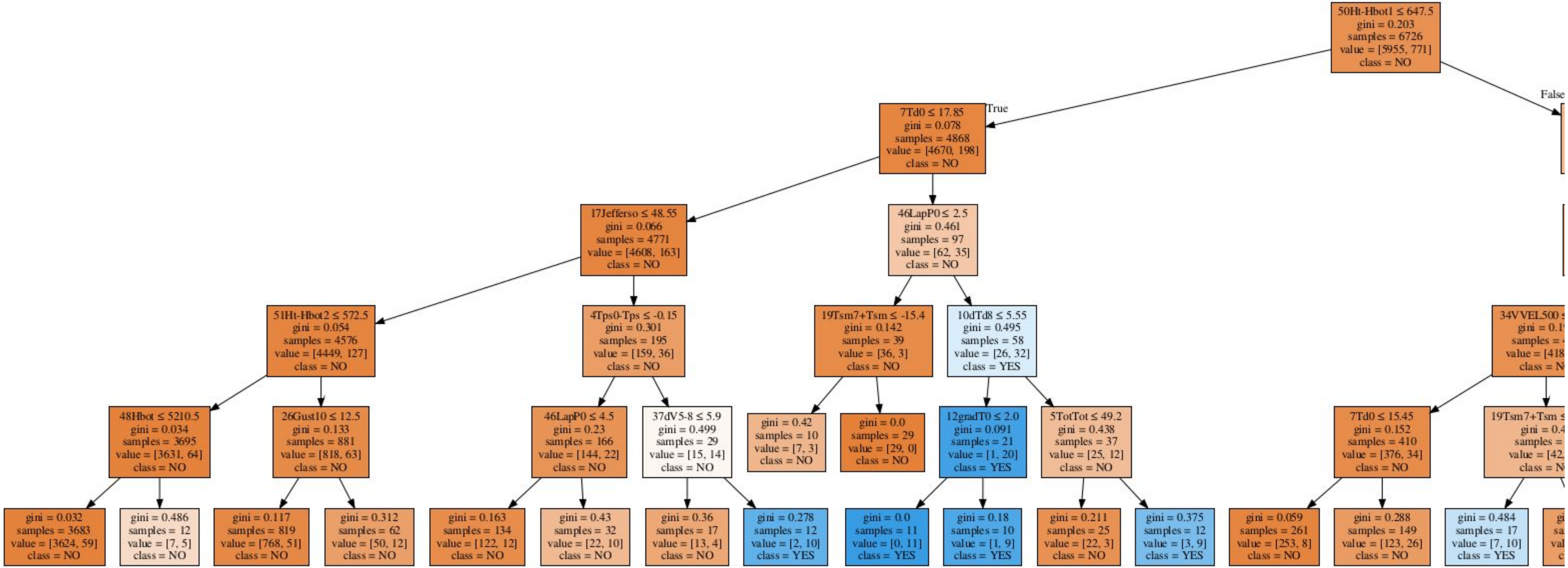
Misses:94   65%

q:0.937 p:0.954 H:0.934

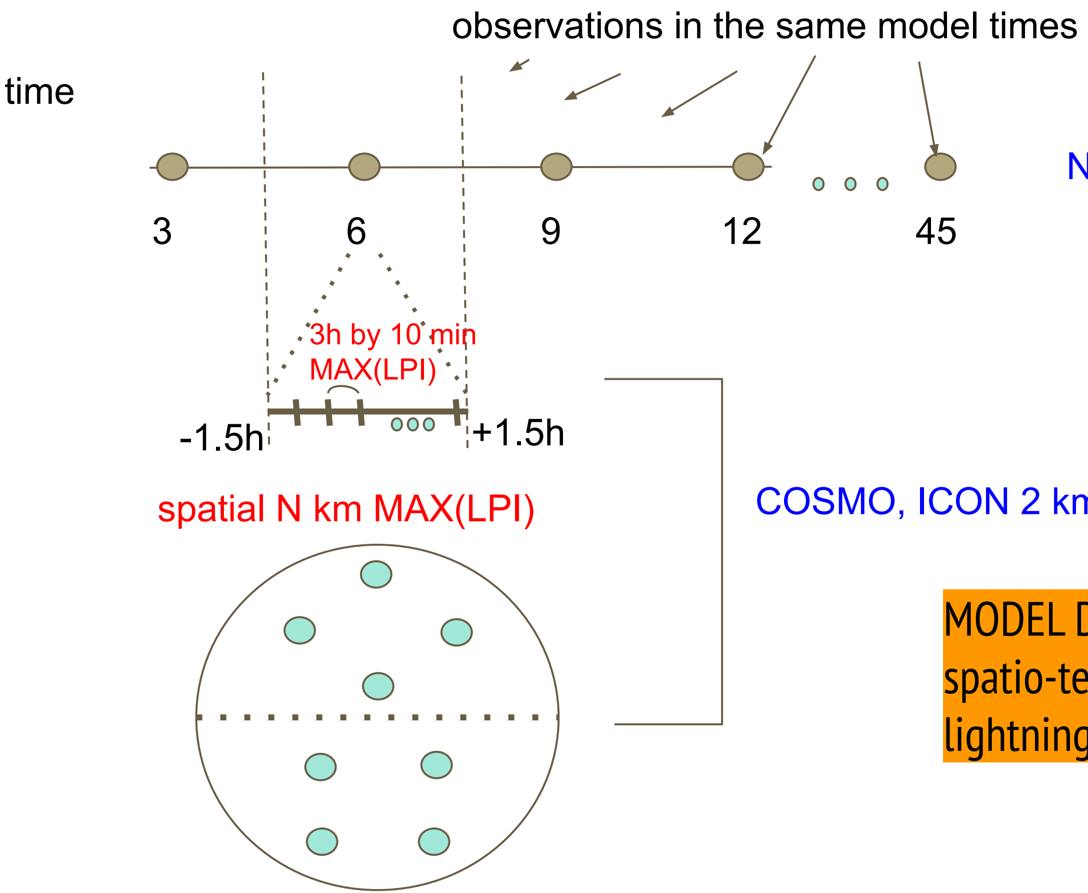
2014-2021 limited cluster tree example (right part)



5-depth tree example (left part)



Comparison of the model and postprocessing data



NN, ML (based on COSMO 13 km)

COSMO, ICON 2 km aggregation

MODEL DATA PROBLEM: optimization of the spatio-temporal aggregation of the model lightning output is required

Aug 2021

MODEL
total synop:
45063
lightening
events:
2849

ML/NN
total synop:
22507
lightening
events:
1423

spatial aggregation (km)	13	20	50	100	point	point
COSMO					ML(0.3B)	ML(NB)
total	45045	45045	47025	47025	22507	22507
hits “yes”	129 (4%)	220 (7%)	663 (23%)	1208 (42%)	548 (40%)	249 (17%)
misses “yes”	2665 (93%)	2574 (90%)	2186 (76%)	1641 (57%)	875 (61%)	1174 (83%)
false alarm “yes”	198 (6%)	361 (12%)	1354 (47%)	3639 (127%)	4072 (286%)	1591 (111%)
ICON					NN(0.3B)	NN(NB)
total	45045	45045	47025	47025	22507	22507
hits yes “yes”	181 (6%)	295 (10%)	905 (31%)	1590 (55%)	576 (40%)	14 (1%)
misses “yes”	2613 (91%)	2499 (87%)	1944 (68%)	1259 (44%)	847 (59%)	1409 (99%)
false alarm “yes”	239 (8%)	416 (14%)	1670 (58%)	4399 (154%)	2951 (207%)	155 (11%)

The conclusions

Machine Learning tech require thoroughly training database preparing. The quality of training dataset affects dramatically. Investigation for developing common rules of building training datasets required.

Most valuable factors:

1.Balancing cases (yes,no, or classes): optimization task solving required.

2. Historical length of training data set.

3. Time and spatial distributions (geographical clusters, etc).

Both DT and NN could be used and shows similar results. Both approaches could be mixed in ensembles.

Future plans

1. Fight with false alarms.
2. Developing common rules for training dataset building.
3. Building rules for geographical clasterisation.

The further direction of our researching and priority task can be adjusted according to WG4 discussion.