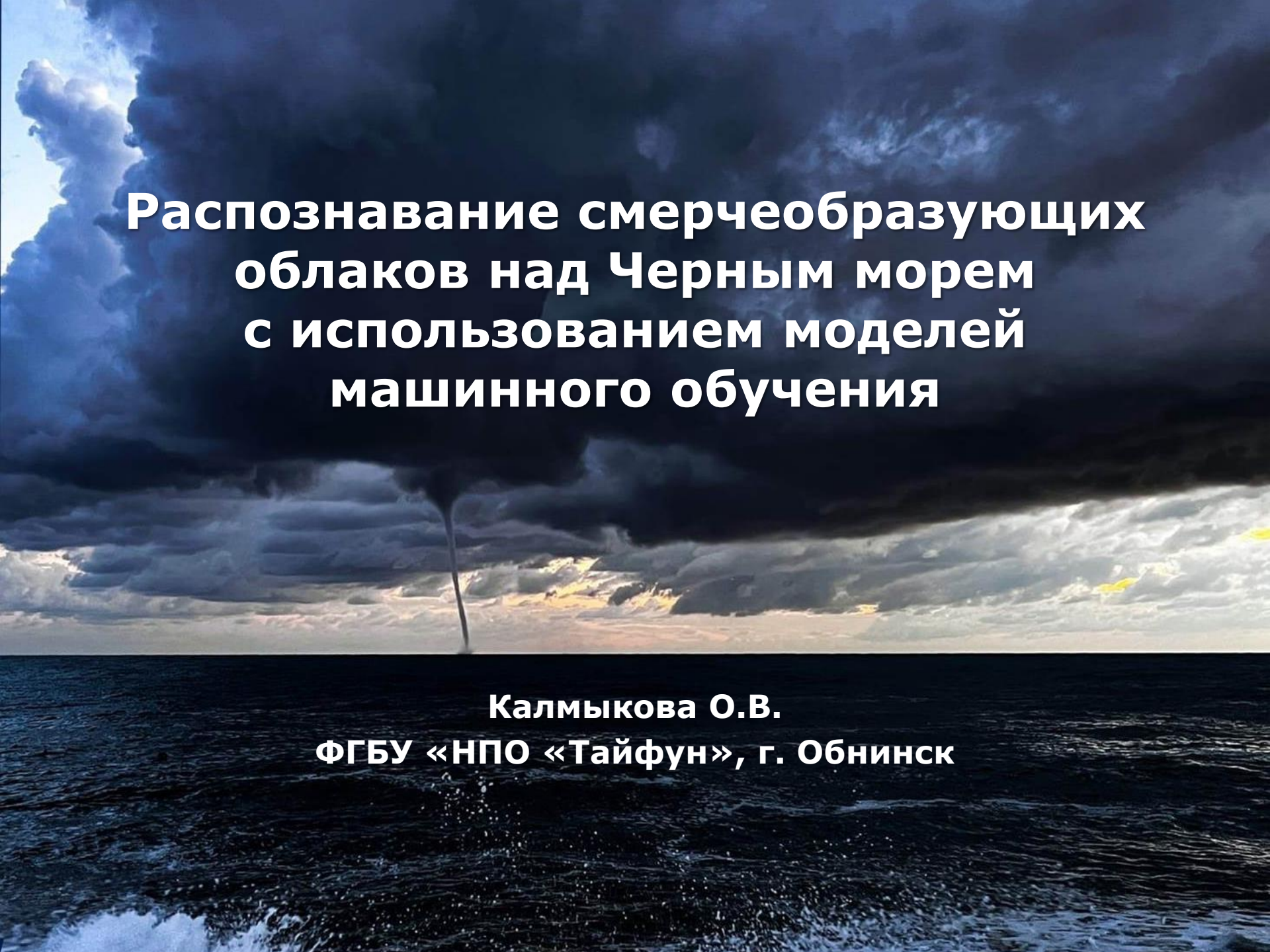


Распознавание смерчеобразующих облаков над Черным морем с использованием моделей машинного обучения

Калмыкова О.В.
ФГБУ «НПО «Тайфун», г. Обнинск

Водяные смерчи над Черным морем

В теплый период года вблизи Черноморского побережья России возникают около 50 смерчей. За последние несколько лет участились случаи их возникновения в опасной близости к побережью.

Год	Кол-во смерчей, вышедших на берег	Кол-во смерчей, близко подошедших к берегу
2020	4	3
2021	3	1
2022	3	0

Последствия выхода смерчей в 2022 г.

29.01.2022 г.
п. Малый Утриш

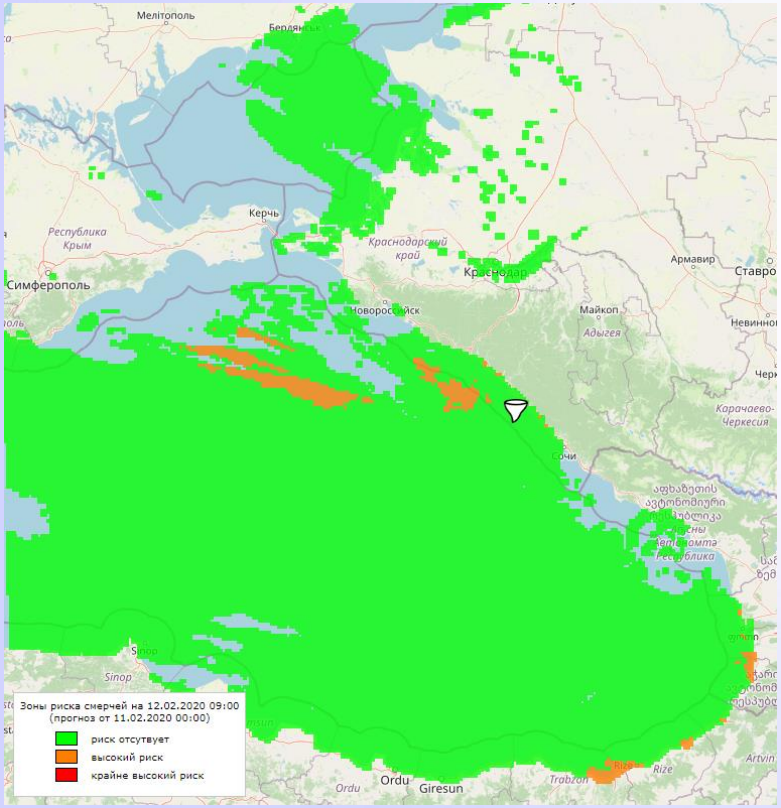
22.06.2022 г.
мкр. Лазаревское

22.07.2022 г. с. Агой

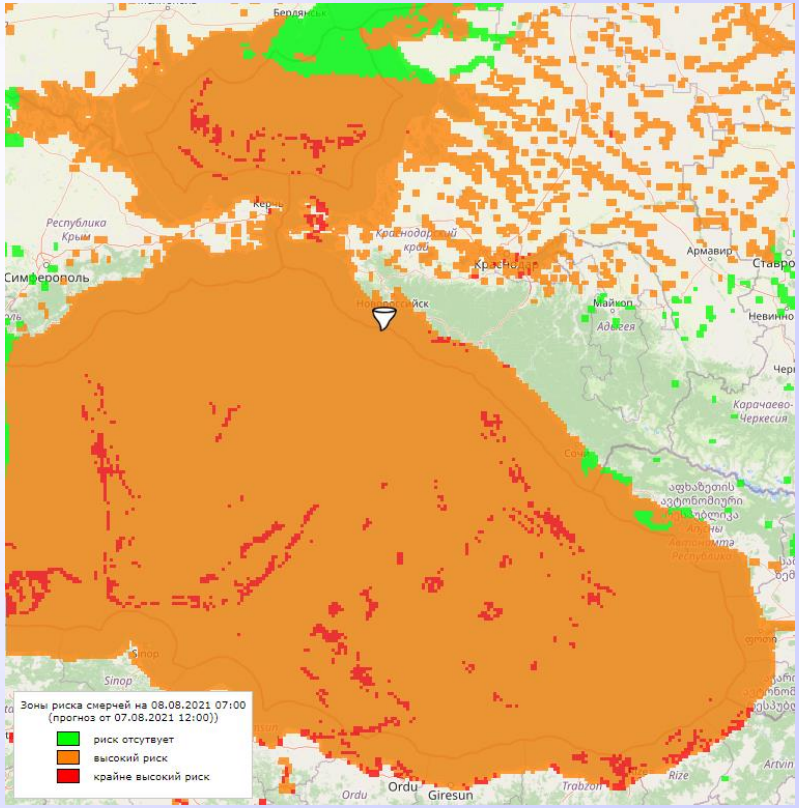
Краткосрочный прогноз черноморских смерчей

Для краткосрочного прогноза черноморских смерчей в теплый период был разработан региональный индекс смерчеопасности WRI. В холодный период года прогноз строится на базе номограммы Szilagyi.

В теплый период предупрежденность смерчей по индексу WRI доходит до 92%. Доля ложных прогнозов может доходить до 72% (известное свойство индексов конвективной неустойчивости прогнозировать опасность с определенной долей избыточности). Индекс WRI может быть использован как первое приближение при построении прогноза смерчей.

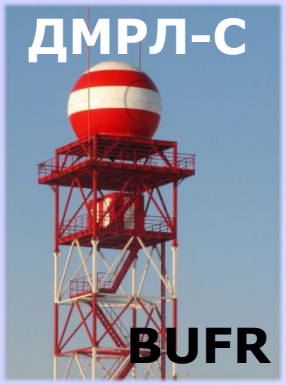


Холодный период (прогноз по номограмме)



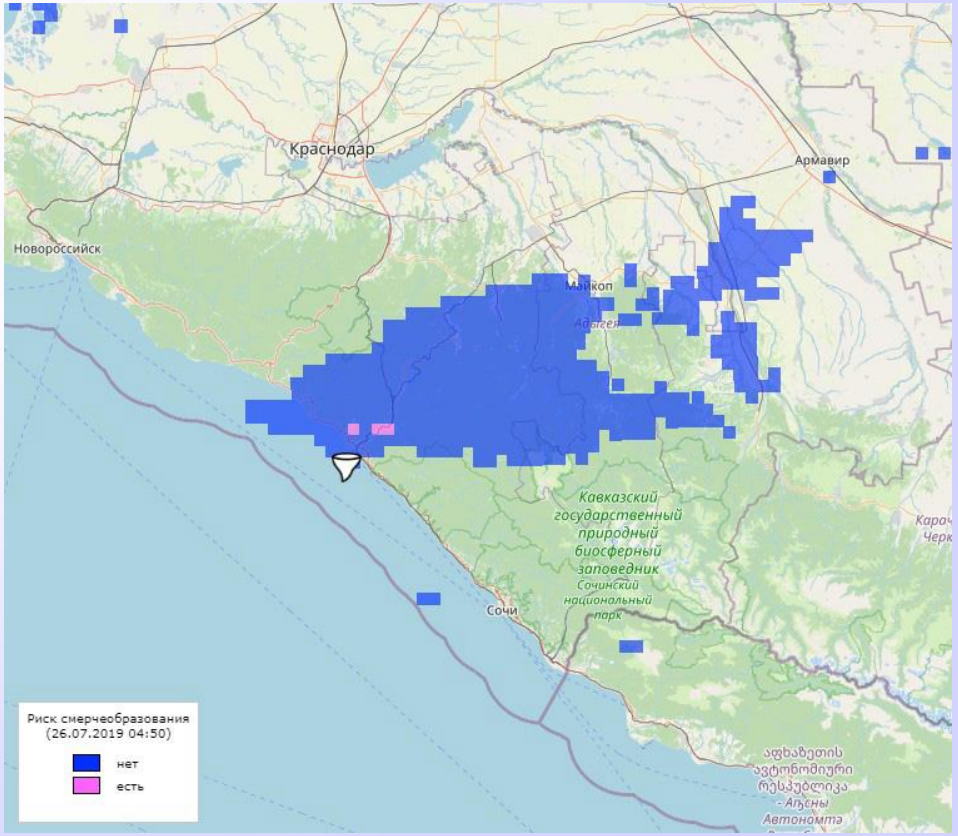
Теплый период (прогноз по индексу WRI)

Наукастинг черноморских смерчей по радиолокационным данным (распознавание смерчеобразующих облаков)



$$R = \begin{cases} dBZ_m \geq 40\ dBZ \\ H_{ВГО} \geq 10\ км \\ VIL \geq 1\ кг/м^2 \\ C_Я \in [осадки, ливень, гроза] \end{cases}$$

Карта риска смерчеобразования



Недостатки алгоритма распознавания смерчеобразующих облаков

- В основе работы алгоритма лежит простейший пороговый принцип выявления областей с высокими значениями радиолокационных характеристик, однако не все смерчеобразующие облака по своим характеристикам достигают используемых в алгоритме пороговых значений. Как следствие, низкая точность распознавания - от 48 до 66% (по данным за 2019-2021 г.).
- Не учитывается специфика смерчеобразующих облаков - диагностируются конвективные системы с возможными опасными явлениями, в том числе смерчами. Как следствие, алгоритм показывает высокую долю ложных тревог.

Новый алгоритм на базе машинного обучения

При разработке нового алгоритма был использован подход к машинному обучению, который предусматривает несколько этапов работы:

- постановка задачи – в данном случае решалась задача бинарной классификации облаков (смерчеобразующие облака (класс W) и облака без смерчей (класс NW))
- описание объектов – формирование признакового пространства
- разметка объектов – отнесение их к определенному классу
- предобработка данных – оценка распределения признаков, фильтрация выбросов, оценка степени корреляции признаков, их нормализация
- выбор модели и параметров обучения
- обучение модели
- тестирование модели

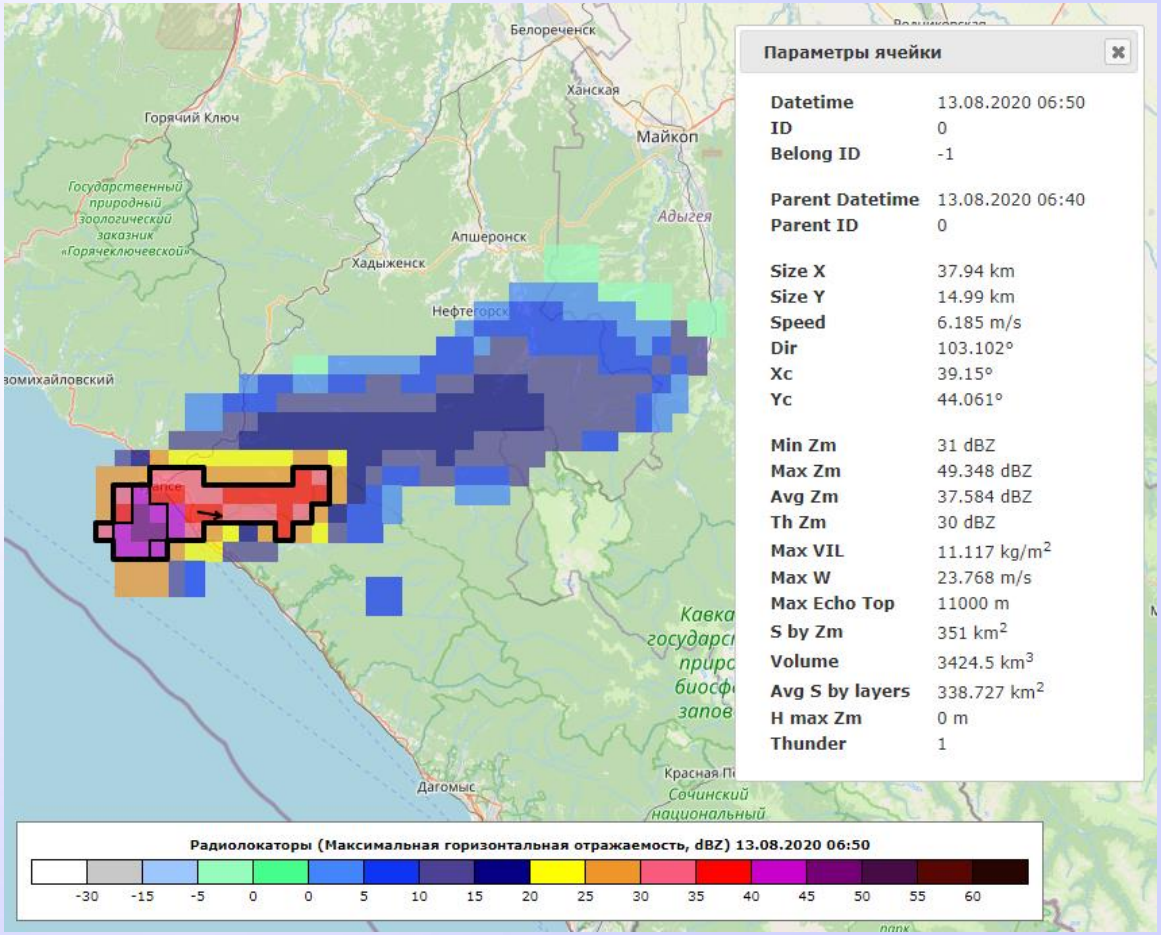
scikit learn
machine learning in Python

Описание объектов

Для описания конвективных систем со смерчами и без смерчей была разработана схема идентификации и последующего прослеживания соответствующих систем по данным о радиолокационной отражаемости. По результатам работы данной схемы формируется набор объектов с соответствующими радиолокационными характеристиками.

Выделение границ конвективных систем осуществляется по пороговому значению в 30 dBZ.
В границах найденных систем осуществляется поиск зон активной конвекции (ЗАК) с запредельными значениями радиолокационной отражаемости.

Схема прослеживания опирается на анализ коэффициента корреляции между объектами по значениям максимальной отражаемости, кроме того учитывается степень подобия объектов по объему и в пространстве.



Выбор параметров (признаков) моделей классификации

Среди множества радиолокационных характеристик конвективных систем для использования в моделях классификации были выбраны 26 параметров, которые в зависимости от алгоритма их расчета могут быть разделены на четыре группы

Группа	Список признаков
мгновенные значения радиолокационных характеристик	1) максимальная горизонтальная отражаемость (max_zm) 2) максимальная вертикально интегрированная водность (max_vil) 3) максимальная высота верхней границы облачности (max_h) 4) максимальная скорость конвективного потока (max_w) 5) объем облака (volume) 6) отметка грозоактивности облака (is_thunder) 7) площадь сечения облака на максимальном уровне положительных значений горизонтальной отражаемости (max_pos_z_s) 8) высота столба с большими значениями дифференциальной отражаемости (> 5 dB) (column_depth) 9) разность максимального уровня положительных значений горизонтальной отражаемости и уровня верхней границы столба дифференциальной отражаемости (dif_h-col2)
скорости изменения значений радиолокационных характеристик	1) скорость изменения max_zm (ds_max_zm) 2) скорость изменения max_vil (ds_max_vil) 3) скорость изменения max_h (ds_max_h) 4) скорость изменения max_w (ds_max_w) 5) скорость изменения volume (ds_volume)
максимальные значения радиолокационных характеристик за последние 60 минут	1) максимальное значение max_zm (max_max_zm) 2) максимальное значение max_vil (max_max_vil) 3) максимальное значение max_h (max_max_h) 4) максимальное значение max_w (max_max_w) 5) максимальное значение volume (max_volume) 6) частота диагностирования грозовых разрядов в облаке в процентном соотношении (thunder_freq) 7) максимальное число непрерывных шагов роста облака по вертикали (max_up_jumps_cnt)
максимальные скорости изменения значений радиолокационных характеристик за последние 60 минут	1) максимальная скорость изменения max_zm (max_ds_max_zm) 2) максимальная скорость изменения max_vil (max_ds_max_vil) 3) максимальная скорость изменения max_h (max_ds_max_h) 4) максимальная скорость изменения max_w (max_ds_max_w) 5) максимальная скорость изменения volume (max_ds_volume)

Разметка объектов

На основе данных о смерчах над Черным морем за период с 2019 по 2021 гг. был сформирован набор из 80 объектов класса W (смерчеобразующие облака). Учитывались случаи регистрации смерчей с известным временем их возникновения.

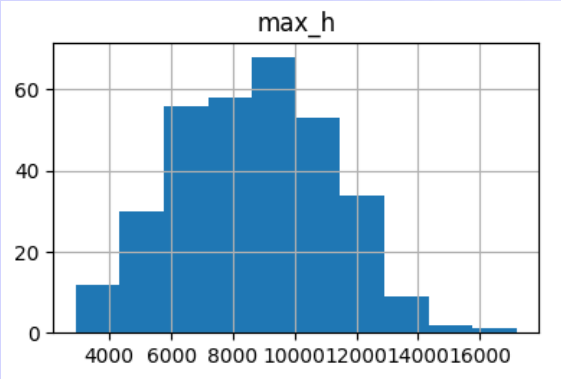
	max_zm	max_vil	max_h	max_w	volume	is_thunder	max_pos_z_s	column_depth	dif_h-col2	ds_max_zm	ds_max_vil	ds_max_h
count	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000
mean	45.400291	7.253678	9916.404506	21.548666	4159.856250	0.725000	281.925000	2.187500	1.312500	0.253779	0.167578	47.271028
std	5.070172	6.316094	2809.283395	5.363925	5461.431756	0.449331	444.700433	1.332113	3.091817	0.385284	0.322140	97.874603
min	31.239442	0.643662	3300.000000	11.183330	189.000000	0.000000	9.000000	0.000000	0.000000	-0.300000	-0.366105	-147.864848
25%	41.883675	3.044059	8316.957977	17.278340	1193.625000	0.000000	60.750000	2.000000	0.000000	0.000000	0.037895	-11.335537
50%	46.007974	6.190857	9986.052724	21.474578	1971.000000	1.000000	112.500000	2.000000	0.000000	0.206773	0.128050	27.121294
75%	49.496626	8.745195	11810.646146	24.569208	3693.375000	1.000000	270.000000	3.000000	0.250000	0.361704	0.223852	105.280683
max	53.946654	44.341057	17200.000000	35.629528	23764.500000	1.000000	2403.000000	8.000000	11.000000	1.889015	2.344394	310.000000

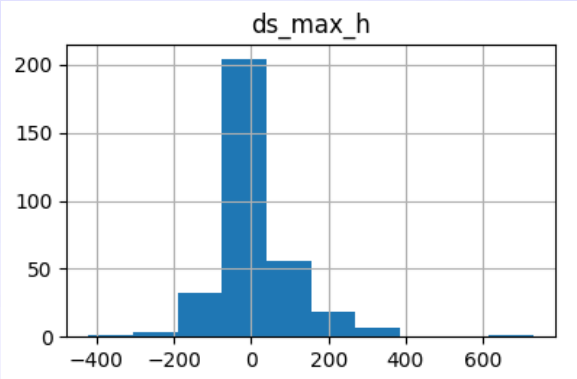
За тот же период случайным образом были отобраны данные класса NW об облаках без смерчей. Для имитации реальной ситуации объектов класса NW было сформировано заведомо большое, чем объектов класса W.

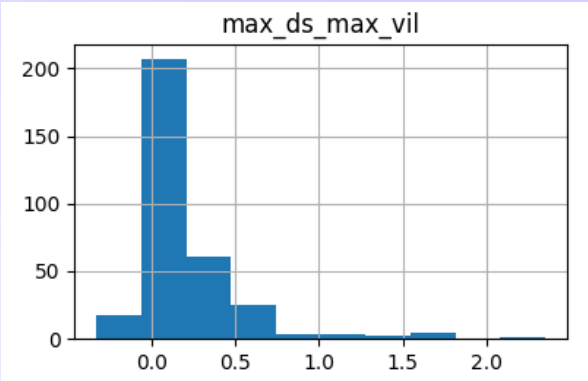
	max_zm	max_vil	max_h	max_w	volume	is_thunder	max_pos_z_s	column_depth	dif_h-col2	ds_max_zm	ds_max_vil	ds_
count	243.000000	243.000000	243.000000	243.000000	243.000000	243.000000	243.000000	243.000000	243.000000	2.430000e+02	243.000000	243
mean	38.027760	3.132585	8584.092367	17.140312	2844.092593	0.362140	173.444444	1.160494	5.234568	6.976615e-03	0.007911	8
std	5.946849	4.966496	2178.145051	4.171162	9277.712330	0.481611	676.516167	1.549114	4.268678	4.582751e-01	0.387439	99
min	30.000000	0.386574	3300.000000	10.682984	81.000000	0.000000	9.000000	0.000000	0.000000	-1.412342e+00	-0.909494	-229
25%	33.714889	1.002581	7060.143793	14.348292	378.000000	0.000000	27.000000	0.000000	0.000000	-2.331718e-01	-0.069927	-45
50%	36.607619	1.692282	8652.213625	16.281843	886.500000	0.000000	63.000000	0.000000	7.000000	-3.552714e-16	-0.003649	-0
75%	41.754632	3.106514	10003.042970	18.878854	2040.750000	1.000000	108.000000	2.000000	9.000000	1.806482e-01	0.039481	39
max	62.000000	54.984781	14564.373673	33.810088	118539.000000	1.000000	8955.000000	9.000000	11.000000	2.359608e+00	4.183176	390

Предобработка данных

Предварительная обработка данных показала, что большинство признаков имело нормальное распределение или близкое к нему.



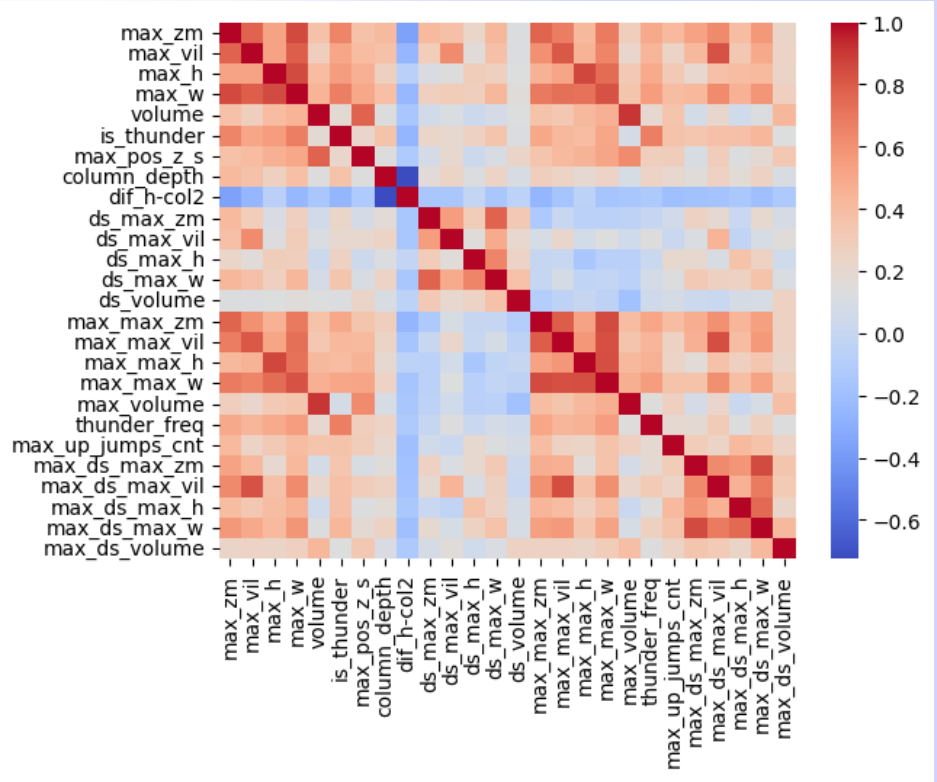




Между некоторыми признаками (7 пар) была выявлена высокая степень положительной корреляции, однако было принято решение не исключать коррелирующие признаки из моделей.

Проведена стандартизация пространства признаков (класс StandartScaler).

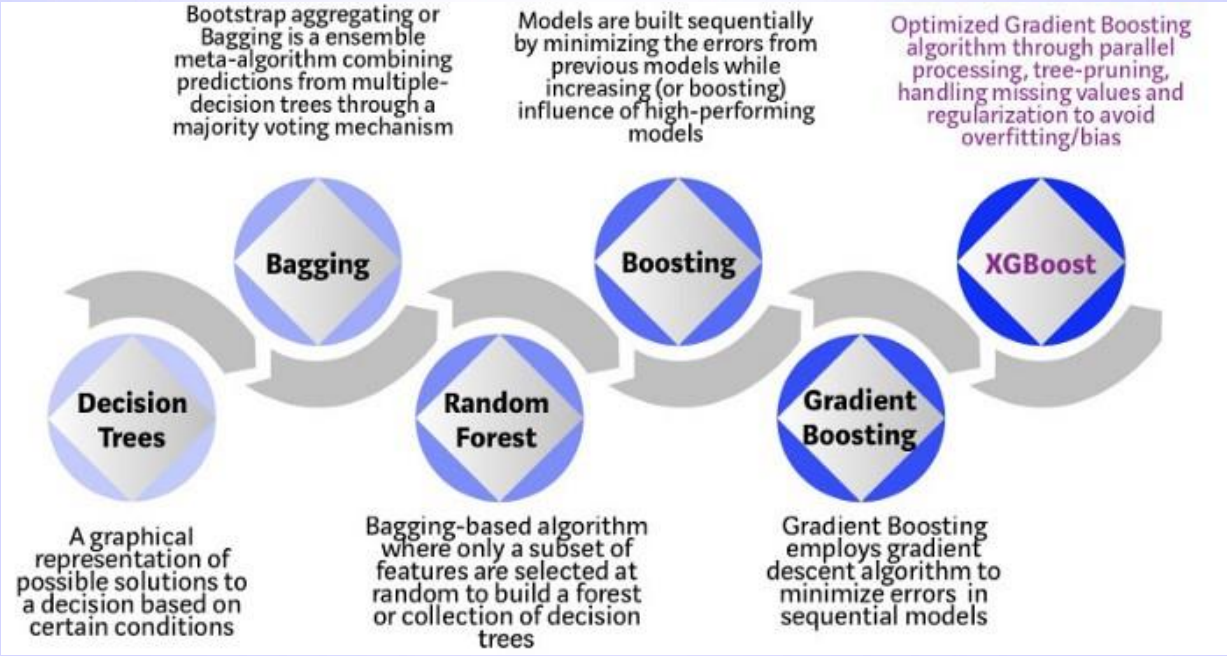
Перед обучением совокупная выборка данных об объектах класса W и NW была разделена на два подмножества: обучающее (70%) и тестовое (30%). Методом бутстрэпа размер обучающей выборки был увеличен в 5 раз.



Выбор моделей

Для решения задачи классификации на исходном наборе данных было проведено предварительное тестирование следующих моделей машинного обучения:

- **дерево решений DecisionTreeClassifier (DT)**
- **случайный лес RandomForestClassifier (RF)**
- **логистическая регрессия LogisticRegression (LR)**
- стохастический градиентный спуск SGDClassifier
- модель k-ближайших соседей KNeighborsClassifier
- модель опорных векторов SVC
- нейронная сеть (многослойный перцептрон) MLPClassifier
- модель адаптивного бустинга AdaBoostClassifier
- модель градиентного бустинга GradientBoostingClassifier
- модель оптимизированного градиентного бустинга XGBClassifier



Обучение моделей

С отобранными моделями были проведены вычислительные эксперименты (500 итераций), в ходе которых случайным образом производилось разбиение исходного набора данных на тестовую и обучающую выборки (аналог кросс-валидации с возможностью пересечения групп). После этого проводилось повторное обучение моделей и оценивалось качество их работы. Критерием выбора наилучших конфигураций моделей стал показатель предупрежденности смерчей (POD) - то насколько хорошо построенные модели распознавали смерчеобразующие облака.

Модель DT	Класс W	Класс NW
Прогноз класс W	20	7
Прогноз класс NW	4	66

POD = 83%

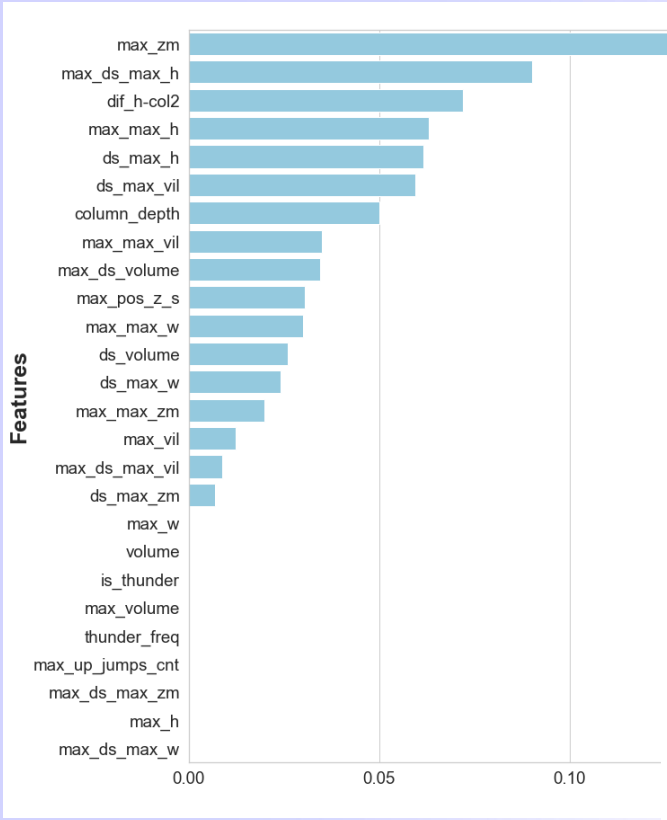
Модель RF	Класс W	Класс NW
Прогноз класс W	21	8
Прогноз класс NW	3	65

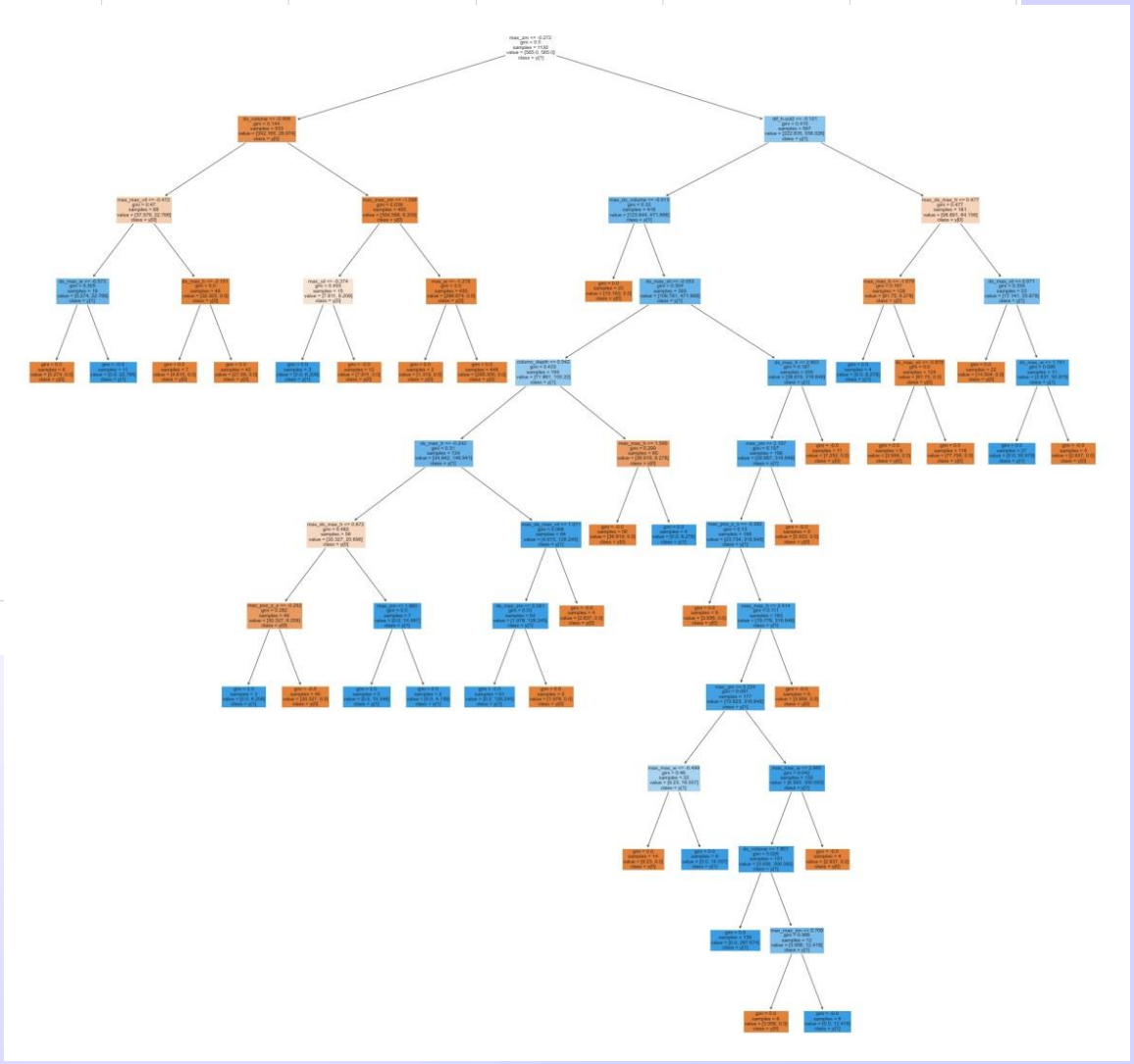
POD = 88%

Модель LR	Класс W	Класс NW
Прогноз класс W	24	17
Прогноз класс NW	0	56

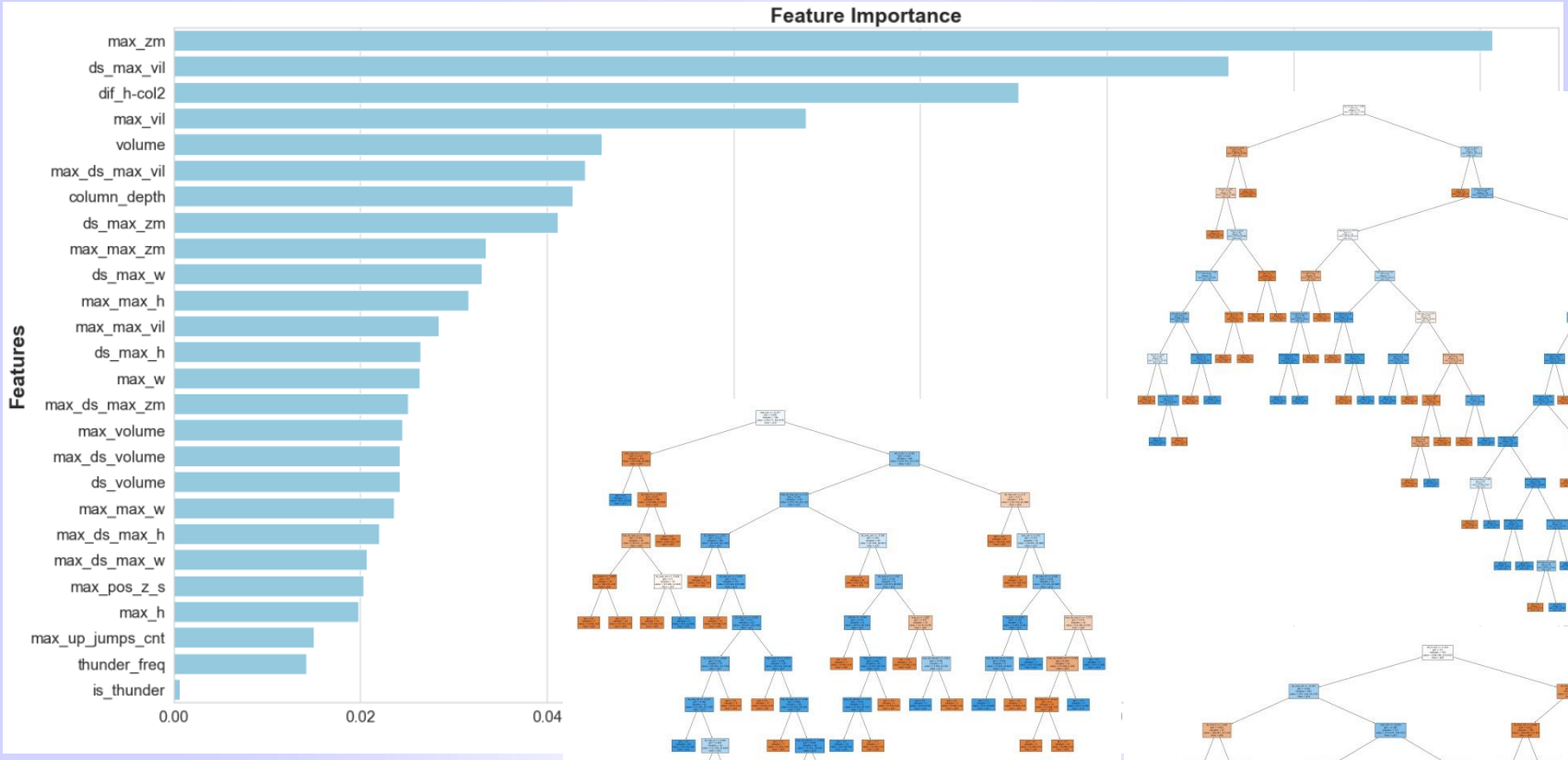
POD = 100%

Модель Decision Tree





Модель Random Forest (100 случайных деревьев)



Модель Logistic Regression (бинарная классификация)

$$z = \sum_{i=0}^{m} x_i \cdot w_i$$

$$\downarrow$$

$$p = \frac{1}{1 + e^{-z}}$$

$$\downarrow$$

$$y = \begin{cases} 1, p \geq 0.5 \\ 0, p < 0.5 \end{cases}$$

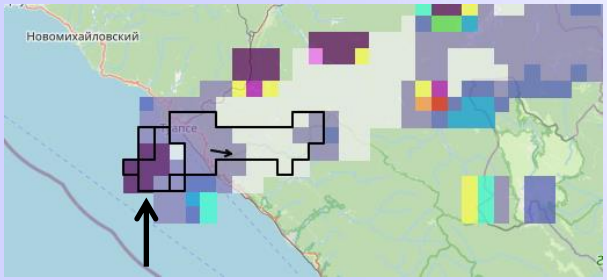
x – множество признаков модели
w – веса признаков

	w
max_ds_max_vil	1.768934
max_w	1.361512
volume	1.025060
max_zm	0.903461
max_max_zm	0.789049
ds_max_zm	0.592038
max_ds_max_h	0.532926
max_pos_z_s	0.506876
max_h	0.380363
max_up_jumps_cnt	0.230841
max_vil	0.127238
max_ds_max_zm	0.043332
thunder_freq	0.010667
ds_max_w	0.001355
is_thunder	-0.074491
ds_volume	-0.117894
ds_max_h	-0.170889
max_max_h	-0.442514
column_depth	-0.506043
max_max_w	-0.525856
max_ds_volume	-0.693945
max_ds_max_w	-0.936050
max_volume	-1.231515
ds_max_vil	-1.333207
dif_h-col2	-1.547351
max_max_vil	-2.020894

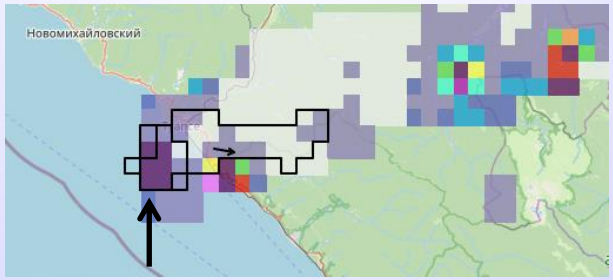
Физическая интерпретация моделей

- Важнейшим критерием разграничения смерчеобразующих облаков и облаков без смерчей является мгновенное значение максимальной отражаемости – с увеличением отражаемости увеличивается вероятность возникновения смерчей.
- По скоростям изменения высоты верхней границы облачности и вертикально интегрированной водности можно диагностировать быстроразвивающиеся облака, которые с большой долей вероятности могут быть связаны со смерчами.
- В водяных смерчах происходит вынос крупных капель воды в верхнюю часть облака, где они сплющиваются. Локализовать области с «приплюснутой» формой гидрометеоров можно по большим положительным значениям дифференциальной отражаемости (> 5 dB). В случае если подобные значения фиксируются на нескольких подряд идущих вертикальных уровнях, то говорят о наличии соответствующего столба. Признак dif_h-col2 позволяет оценить расположение рассматриваемого столба относительно верхней границы облака – чем меньше значение данного признака, тем больше вероятность того, что в рассматриваемом облаке может или уже формируется смерч.

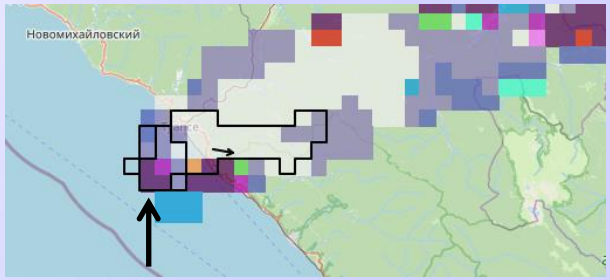
Столб дифференциальной отражаемости с 6 до 9 км



6-7 км



7-8 км



8-9 км

Тестирование моделей

Период тестирования - с июня по сентябрь 2022 г. За это время вблизи побережья Краснодарского края было зарегистрировано всего 26 смерчей (по всей видимости, следствие аномально жаркого лета).

По результатам тестирования было проанализировано качество классификации материнских облаков для 14 смерчей с известным временем их возникновения. Рассматривались только те смерчи, которые достигали поверхности воды. Кроме того оценивалось качество заблаговременного распознавания опасности смерчегенеза.

Предупрежденность смерчей (точность классификации в момент регистрации смерча):

- модель DT – 79%
- модель RF – 93%
- модель LR – 93%
- композиция моделей – 71%
- алгоритм порогового распознавания – 29%

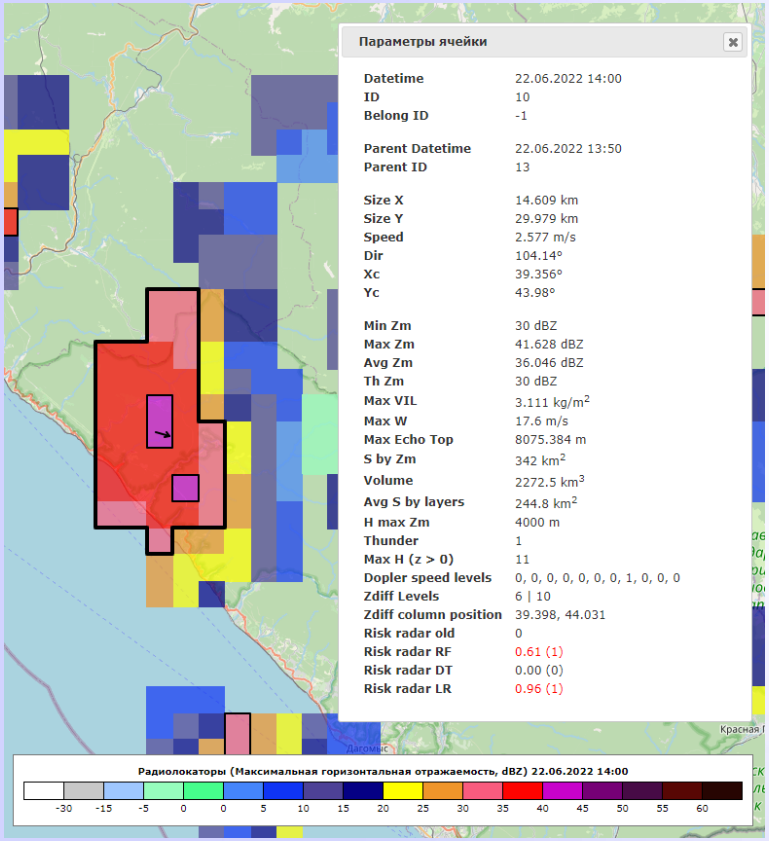
Предупрежденность смерчей (заблаговременный прогноз опасности):

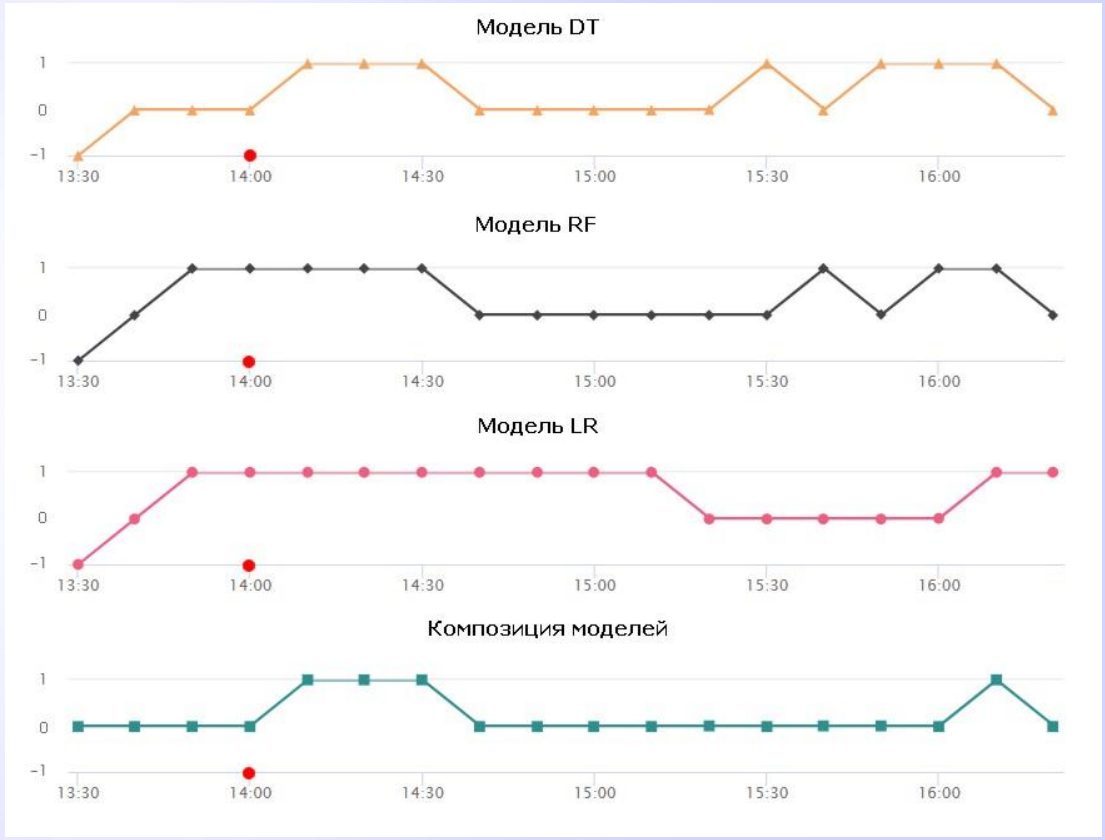
- модель DT – 75%
- модель RF – 83%
- модель LR – 92%
- композиция моделей – 75%
- алгоритм порогового распознавания – 17%

Доля ложных прогнозов:

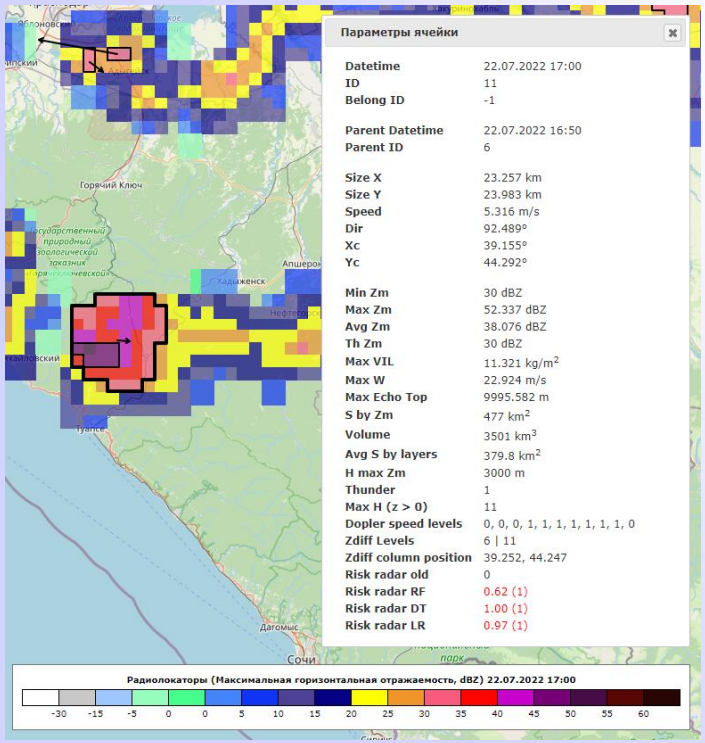
- модель DT – 17%
- модель RF – 12%
- модель LR – 45%
- композиция моделей – 9%
- алгоритм порогового распознавания – 12%

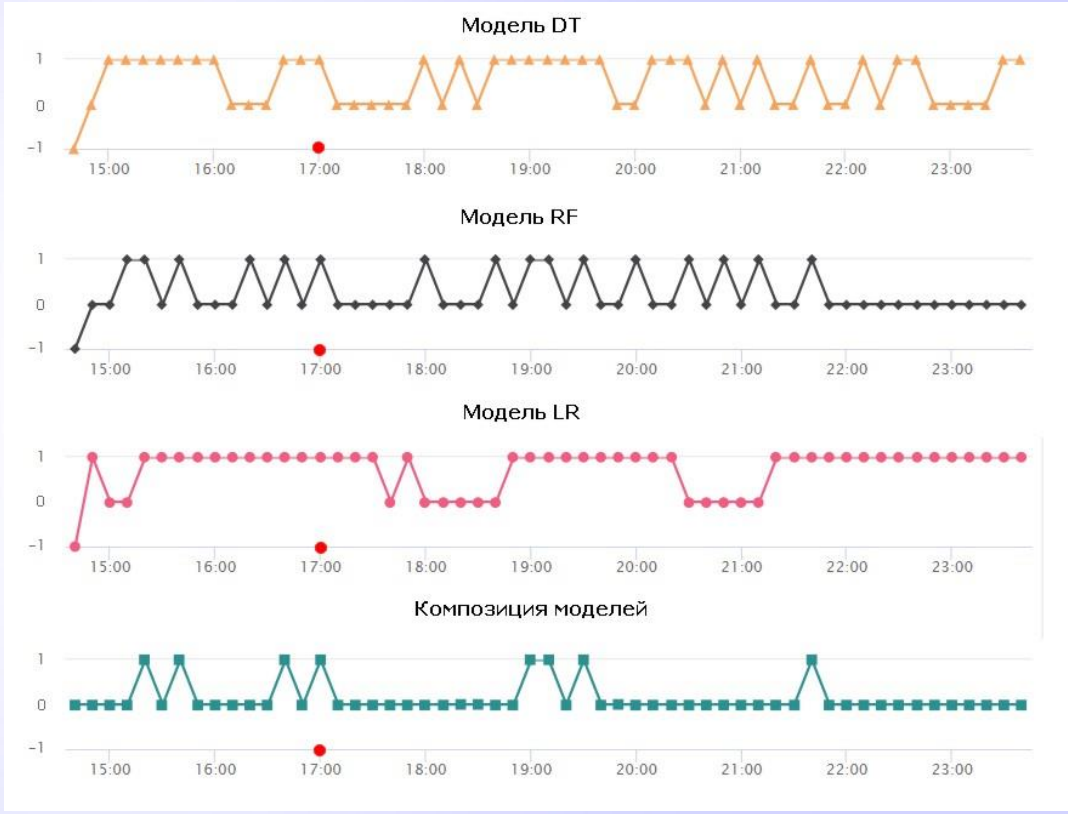
Смерч 22 июня 2022 г. вблизи п. Аше





Смерч 22 июля 2022 г. вблизи п. Агой Туапсинского района





Заключение

- Точность прогноза смерчей может быть существенно повышена за счет использования данных оперативных наблюдений.
- На основе подхода к машинному обучению были построены три модели классификации конвективных облаков, использующие в качестве параметров значения радиолокационных характеристик.
- В ходе непрерывного тестирования модели показали достаточно хорошие результаты по предупрежденности смерчей (до 92%).
- План дальнейшей работы:
  - ✓ накопление новых данных о смерчах, дальнейшее усовершенствование моделей, их тестирование в течение более продолжительного периода времени
  - ✓ учет краткосрочного прогноза смерчеопасности при распознавании смерчеобразующих облаков

Спасибо за внимание!