Inferences on deep convective initiation

Jowan Fromentin

Abstract

We have inferred via linear logistic regression models using Goes satellite channels, an influential domain for deep convective initiation of 10kmx10km over land and 14kmx14km over sea. An influential leading time for deep convective initiation was estimated to be 15 minutes. Cooling of the water vapour difference field was found as a mechanism of identification in the model. Via recursive feature elimination Goes satellite channels 7 and 10 were the important channels in identifying deep convective initiation over land and channels 4, 8 and 10 were the most important over sea.

1 Introduction

Convection is the process of converting the potential energy of air parcels into kinetic energy via vertical motion. This occurs as warm air parcels rise to transition into an energetically lower state. Convection becomes 'deep' when the rising air parcel traverses from near the surface to above the 500hPa level. Convection of moist air parcels lead to the initiation of thunderstorms and other severe weather events, meaning understanding of this topic is crucial.

2 data set and Methods

2.1 Tobac flow

Tobac flow [2] is a branch of the Tobac python module for cloud tracking. Tobac flow uses an optical flow algorithm on the brightness temperature field to create a Lagrangian perspective of the cloud. This allows a calculation of the changes of the Water Vapour Difference (WVD) field with time. The WVD field is the difference in water vapour between the upper and lower atmosphere. Tobac Flow also applies thresholding of 0.5 K per minute cooling on the WVD field to define Deep convective initiation (DCI) [1]. Tobac flow thresholding requires that the cooling of the WVD field occurs over at least 3x3 pixels.

This algorithm was applied to GOES 16 satellite data over the gulf of Mexico and its surrounding land masses, as shown in Figure 1.

2.2 Growing cores detected

Once the threshold is passed the initiation cores begin to grow. We define the initiation beginning at the first frame that the threshold is satisfied. We want to consider the conditions before initiation. Each data set contains 8 channels of the GOES 16 satellite shown in Table 1 [6]. Channels coloured in grey in Table 1 were not used due to their lack physical interpretation, making inferences difficult. A domain of 4 frames (including the frame detecting initiation) and 4 pixels either side of the centre pixel of the detected initiation for each channel was selected to create the data set for each initiation, this is visualised in Figure 2. There are 5 minutes between frames and each pixel represents an area of 2kmx2km, meaning our data set spans an area of 18kmx18km over a 20 minute period. In summary each data set consists of an 8x4x9x9 data frame.

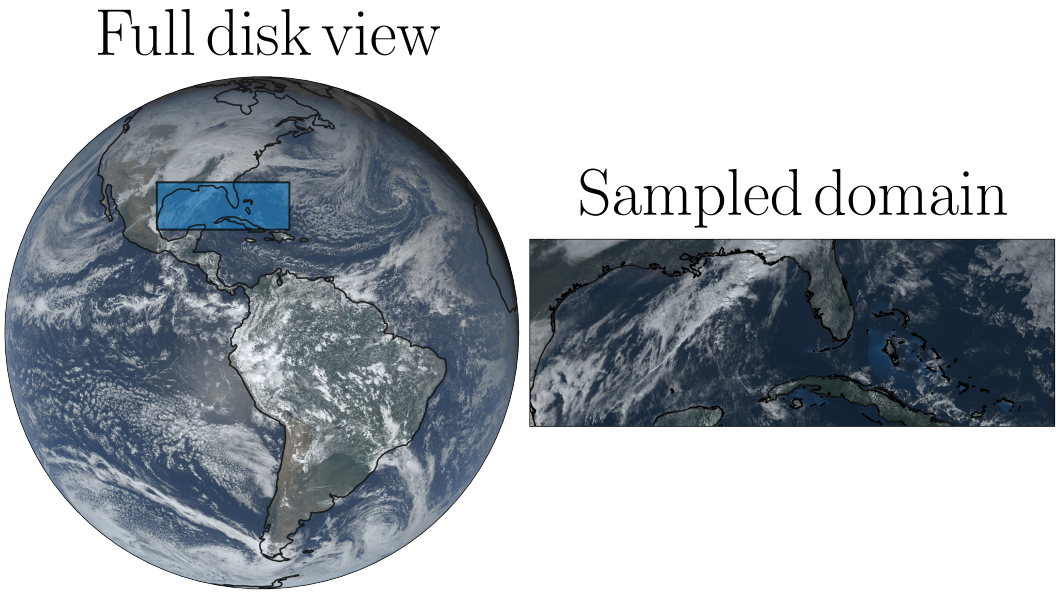

Figure 1: The GOES 16 satellite full disk shown next to the domain that testing and training samples were taken from. Shown in blue on the full disk

Data sets for the negative case (no DCI) were collected from the same coordinates as positive cases.

Table 1: GOES-16 channels. Channels not used coloured in grey.

ABI Band	Central Wavelength (µm)	Type	Nickname
1	0.47	Visible	Blue
2	0.64	Visible	Red
3	0.86	Near-Infrared	Veggie
4	1.37	Near-Infrared	Cirrus
5	1.6	Near-Infrared	Snow/Ice
6	2.2	Near-Infrared	Cloud particle size
7	3.9	Infrared	Shortwave window
8	6.2	Infrared	Upper-level water vapor
9	6.9	Infrared	Midlevel water vapor
10	7.3	Infrared	Lower-level water vapor
11	8.4	Infrared	Cloud-top phase
12	9.6	Infrared	Ozone
13	10.3	Infrared	”Clean” longwave window
14	11.2	Infrared	Longwave window
15	12.3	Infrared	”Dirty” longwave window
16	13.3	Infrared	CO2 longwave

3 Analysis

Convection is initiated via different mechanisms over land and sea. The Sun’s heating of the Earth’s surface causes thermal free convection over land. Whereas convection over sea is caused by cooling higher in the atmosphere creating an unstable air layer above [3]. Meaning that land and sea data sets must be separated for training and testing, as this requires different models to account for these

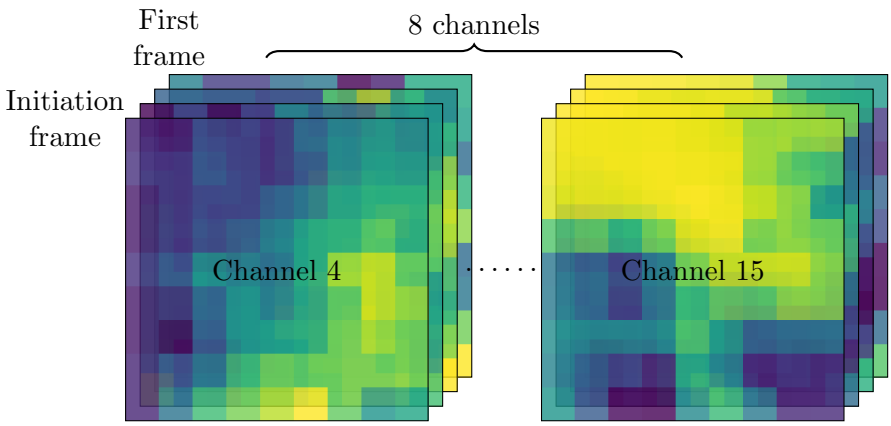

Figure 2: Visualisation of the data sets saved for each instance

mechanisms, as different features and parameters are important to initiation.

Models were trained on the full data set to provide a prediction of whether DCI exists in the final frame. The influence of the leading time and domain on the predictability of DCIs was investigated.

3.1 Leading time

Leading time is the time between DCI and the final frame in the data set. Restricting the time frames the models were trained on changed the model's task from predicting if there was DCI in the final data set frame, to predicting if there was with DCI 5, 10, 15 minsutes leading time. This is visualised in Figure 3. This is to infer the leading time that is influential on DCIs.

3.2 Domain area

An outer ring of pixels was removed from the input domain and the model was rerun until a 1x1 input domain was reached. This is shown in Figure 4 for the 5x5 pixel case. This allows inference on the region that is influential on DCIs.

3.3 Optimization

All models are trained with a spectrum of values for logistic regression hyper-parameter, C [4]. The value of C that maximizes the model accuracy is said to optimise the model. Values of $C < 1$ creates a sparse coefficient domain, meaning many coefficients are zero. Values of $C > 1$ creates a populated coefficient domain, with many low-magnitude, non-zero coefficients. The models described in 3.1 and 3.2 were optimised for every iteration. The model trained on the full data set was optimised.

3.4 Cross validation

Cross validation is used to decouple a model's performance and coefficients from the data set it is trained on, as they depend on the separation of testing and training data. The aim is to infer how a model will generalise to an independent data set, via a k-folding algorithm. The analysis described in Sections 3.1 and 3.2 was cross validated. The optimised model for the full data set was cross validated to find a general distribution for coefficients in each channel.

3.5 Recursive feature elimination

Recursive Feature Elimination (RFE) was performed to determine a ranking for the 'importance' of the channels on the optimised, full data set model. RFE is a common process that repeatedly 'trims' features until a predefined number of features are left. These are the most influential in the model.

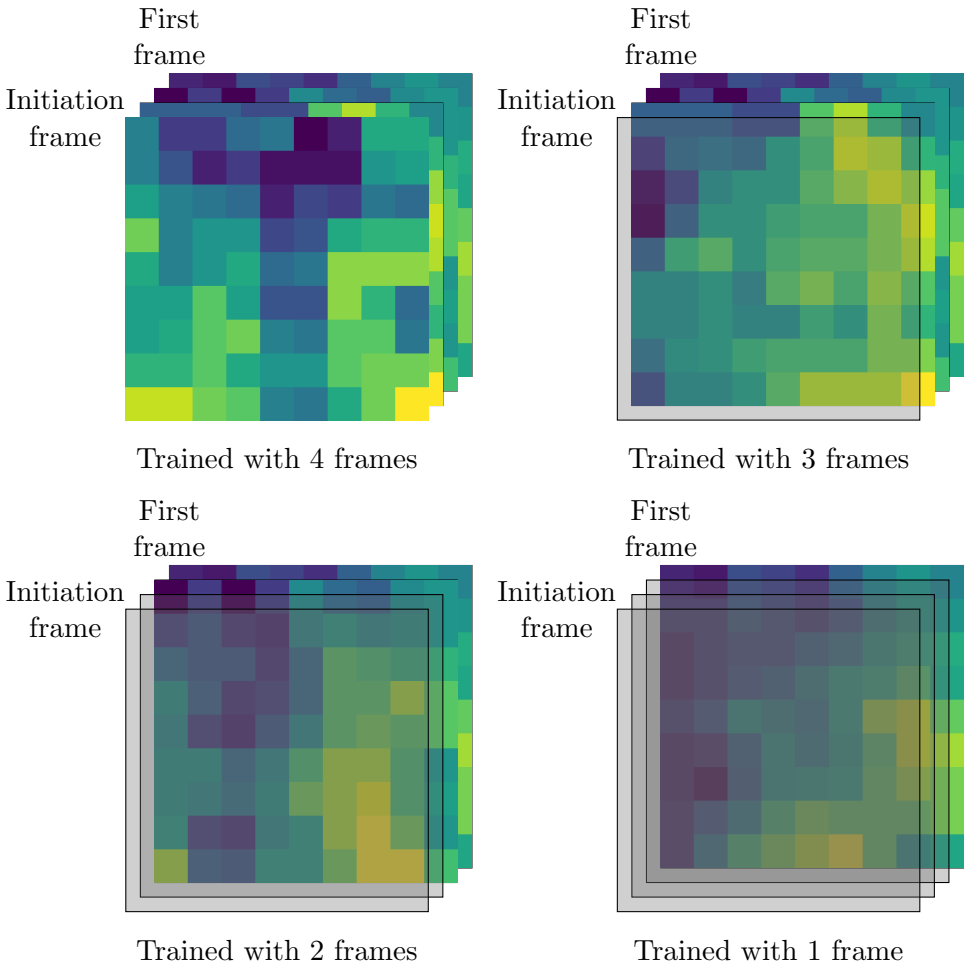

Figure 3: Visualisation of how the leading time changes the training and testing data

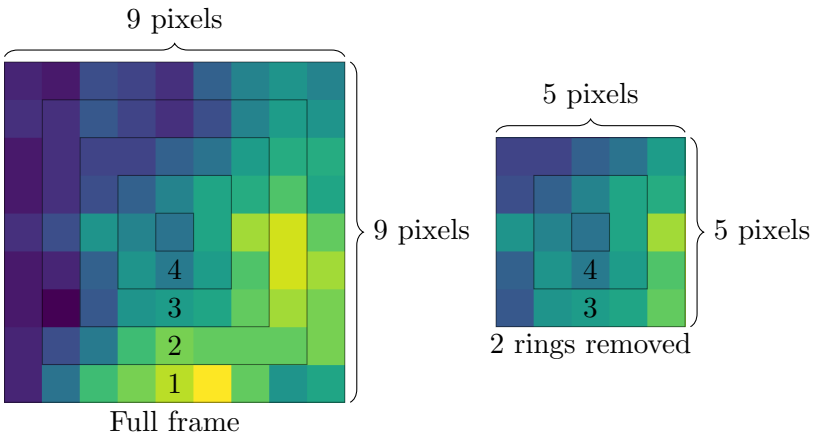

Figure 4: An example of the area domain clipping. This is done for every frame

The RFE applied to the model finds the most influential features (pixels). Channels are ranked by the number of 'important' features in each channel.

4 Results

4.1 Leading time

Figure 5 displays the cross validated model accuracy for land and sea models with leading time. We see a monotonically increasing behaviour for model accuracy with the number of frames the model has been trained on. This is the case for both land and sea models. The land models consistently performed better than sea models. The mean model accuracy values are displayed in Table 2.

Table 2: The mean model accuracy for the cross validated data for changes in leading time

Mean model accuracy				
	Number of frames			
Domain	1	2	3	4
Land	0.691	0.759	0.779	0.801
Sea	0.652	0.727	0.755	0.784

The low accuracy of the 1 frame models implies that conditions 15 minutes before DCI are not consistent. The increase in model accuracy with the addition of a new frame for training is greatest between the 1 and 2 frame models. We can infer that influential conditions leading to DCI are withing the 15 minutes before initiation.

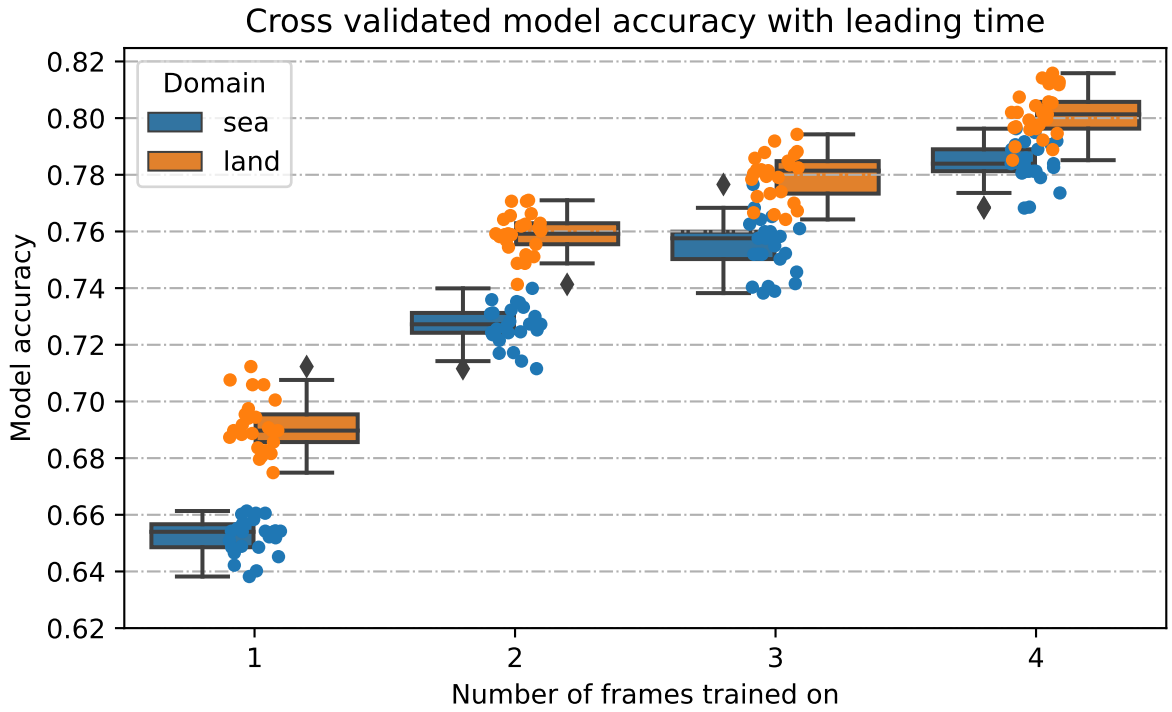

Figure 5: Cross validated model accuracy as the number of frames the model is trained on increases. This has the effect of changing the lead time of the model

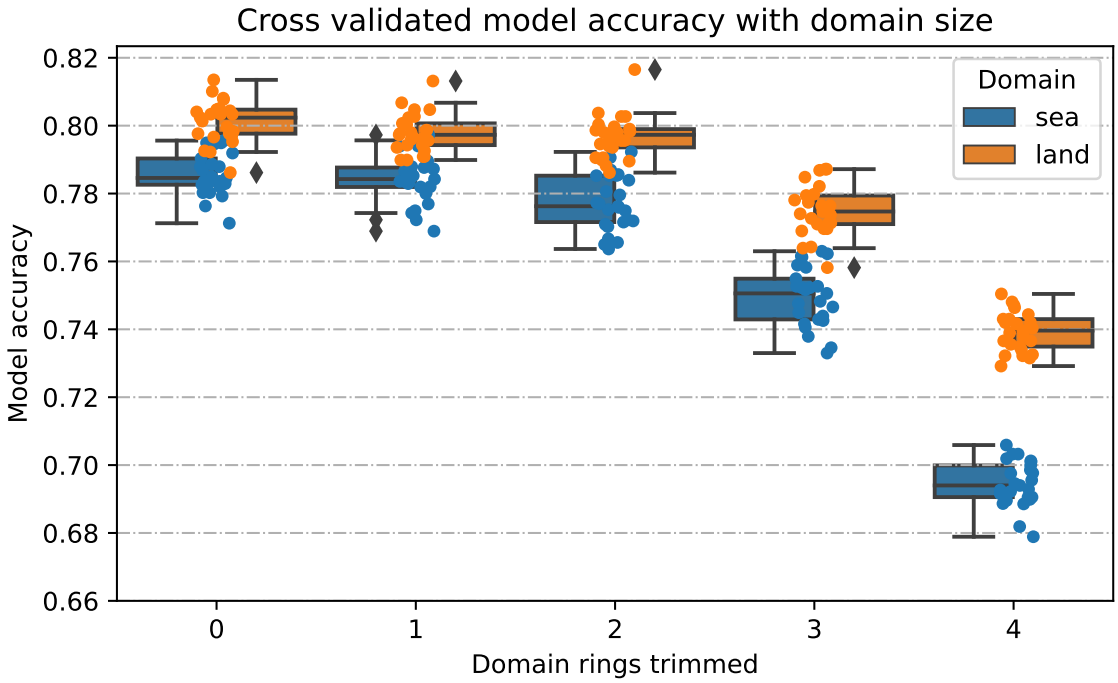

Figure 6: Cross validated model accuracy as the domain the model is trained on is reduced

4.2 Domain area

We see a monotonically decreasing behaviour for model accuracy with the number of rings trimmed. The land model outperformed the sea model for all cases.

The information found in the outermost useful fringe ring is used to identify some DCI cases. When this outer ring is removed the models are less capable of correctly identifying these cases, increasing the variability, and reducing the accuracy for models with that domain.

For the land model there is little difference between the models where 0, 1 or 2 rings have been trimmed. The model where 3 rings have been trimmed has a larger drop in accuracy and an increased variability. This implies that information that is useful for identifying DCI is found in the 1st ring for most cases. For the sea models, we see this large increase in variability when 2 rings are trimmed.

Extending this idea of the increase in variability representing the fringe of the useful domain, we can estimate an 'influential' domain for DCI. Each pixel is roughly 2kmx2km. There is a larger domain over sea that influences DCI than for land. The estimated domains that effect DCI are displayed in Table 3.

Table 3: The estimated domain around the centre of DCI that effects its initiation. The ring number is taken from Figure 4

Estimated influential domain		
	Outermost 'useful' ring number	Most influential domain
Land	3	10kmx10km
Sea	2	14kmx14km

Table 4: Channel ranking and the number of influential features

Channel ranking				
	Sea model		Land model	
Rank	Channel	Important features	Channel	Important features
1	4	17	10	22
2	8	15	7	19
3	10	14	13	12
4	7	9	8	8
5	14	8	4	7
6	15	7	9,14	5
7	9	6		
8	13	5	15	3

Table 5: Performance metrics for the models in the coefficient plots

Domain	Accuracy	F-1 score
Land	0.794	0.771
Sea	0.783	0.764

4.3 Recursive feature analysis

The channel ranking when considering the 81 most influential features (pixels) for land and sea models are shown in Table 4. Figures 7 and 8 display the channel ranking as the number of influential features is increased for land and sea models. The channels change rank throughout, showing that it is difficult to pick the 'best' channel.

However, from Table 4 we can infer that channels 4, 8 and 10 are the most influential for sea models. From Table 1 these represent the 'Cirrus', 'Upper-level water vapour' and 'Lower-level water vapour'. For land models channels 10 and 7 are the most influential. Referring to 'Lower-level water vapour' and the 'shortwave window'.

4.4 Coefficients

4.4.1 Channel plot

The coefficient magnitude can be interpreted as the model's sensitivity to that feature. Features with high coefficient magnitudes are most influential in categorising the data set as a DCI or not.

Figures 9 and 10 display the coefficients for a land model and a sea model respectively. The accuracy and F-1 score of these models are shown in Table 5.

An F-1-score is an alternative to accuracy and is calculated from the harmonic mean of the precision and recall of the model [5].

For both domain models channels 14 and 15 dominated negative and positive coefficients respectively. From Table 1 this implies that low values of longwave and high values of "Dirty" longwave in the build-up are indicative of DCI. Frames that are dominated by positive coefficients followed by a frame of negative coefficients implies that 'cooling' of that channel is influential for DCI or vice versa. This is the case for channel 13 of the land model and channel 10 on a longer time scale. The final frame of channel 13 of the sea model shows negative coefficients in the middle of the frame and positive coefficients around the perimeter. This implies that there are lower values of "Clean" longwave at the centre of DCI.

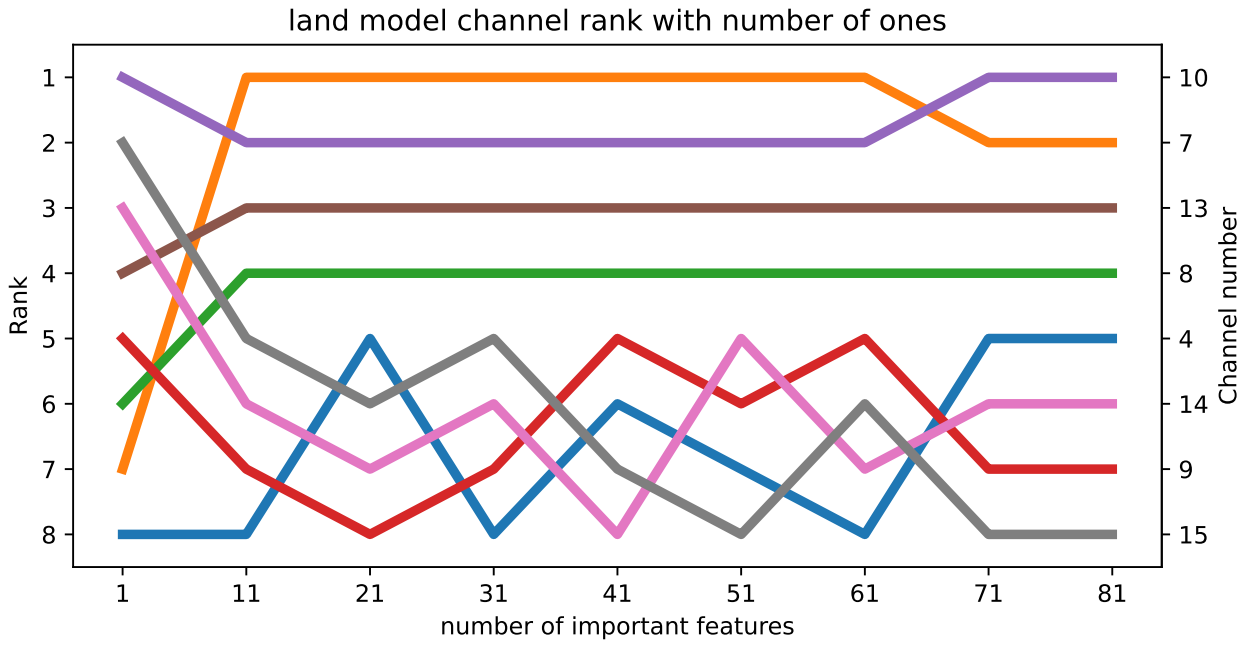

Figure 7: Bump chart for the rank of channels as the number of influential features are increased for land models

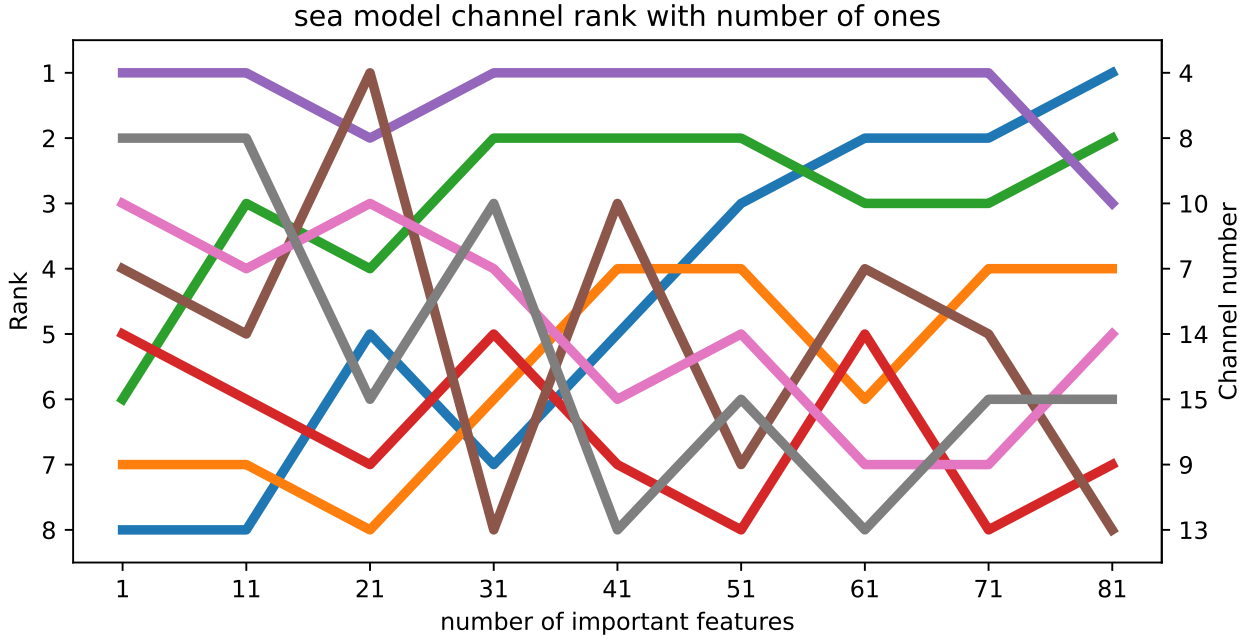

Figure 8: Bump chart for the rank of channels as the number of influential features are increased for sea models

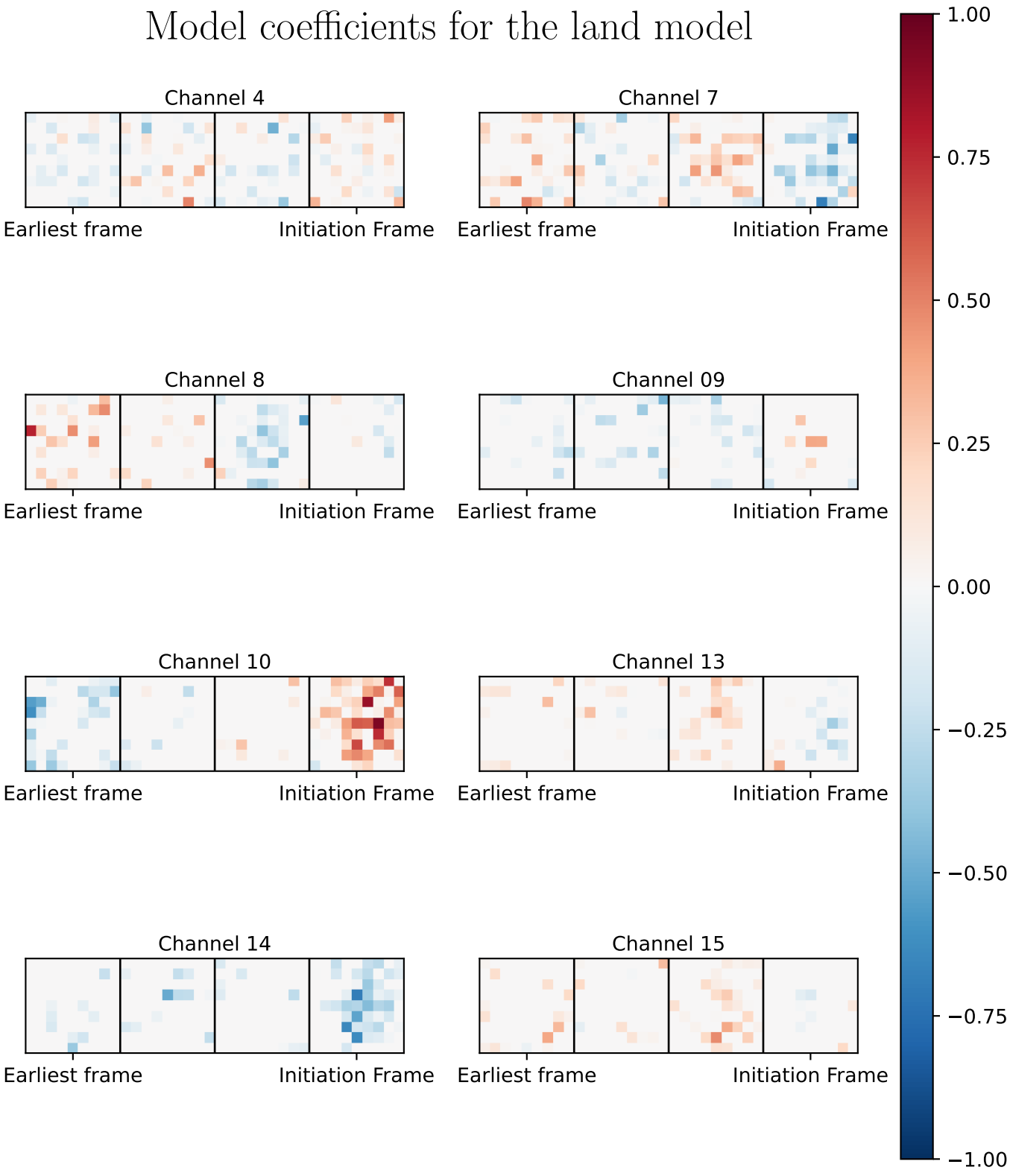

Figure 9: Coefficients for each channel of the land model

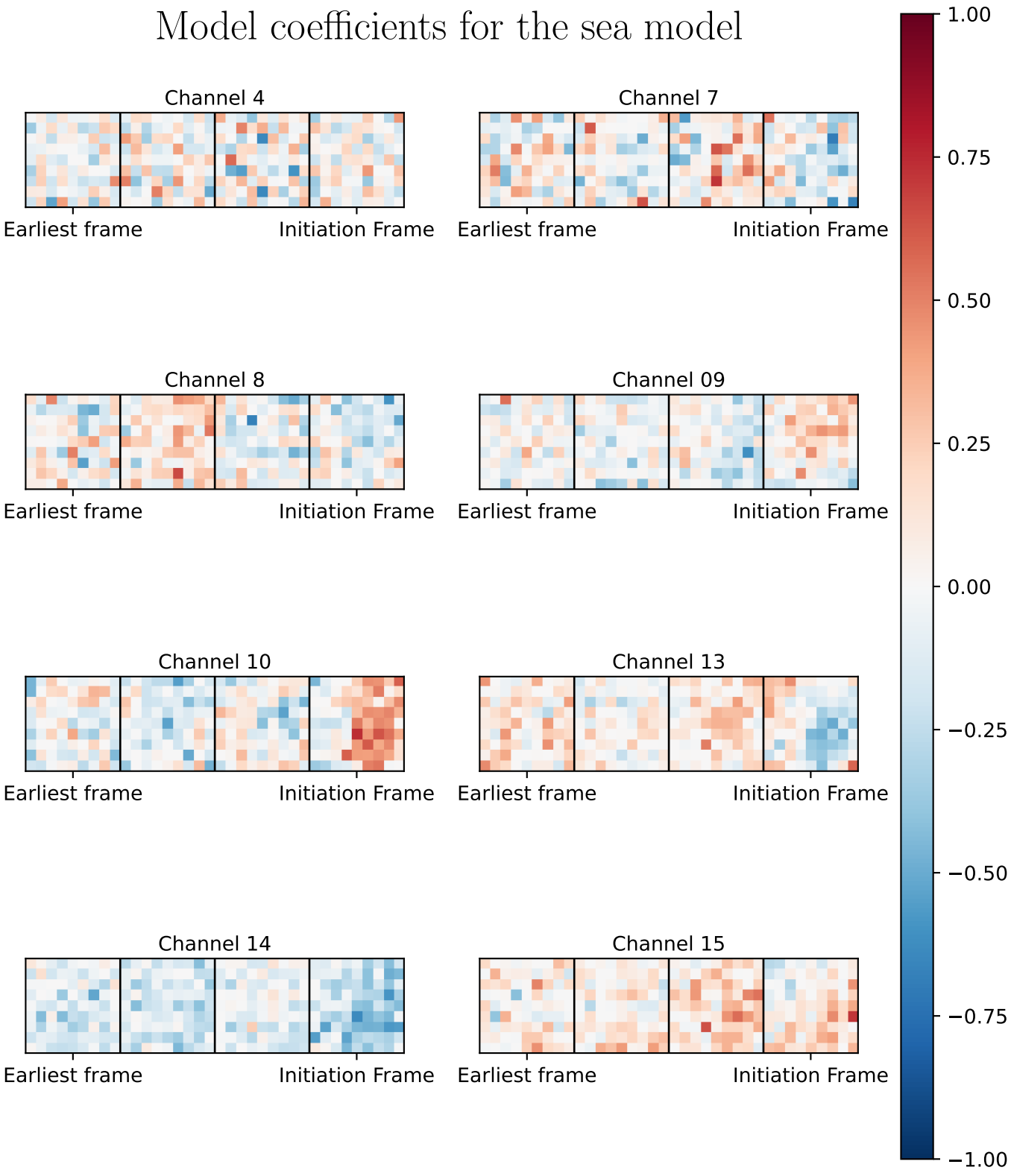

Figure 10: Coefficients for each channel of the sea model

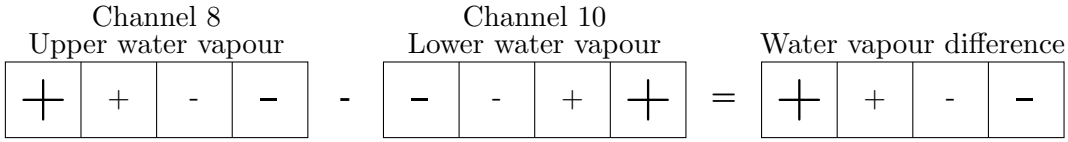

Figure 11: Water vapour difference coefficient construction

In Section 2.1 we defined WVD as:

$$\text{Water vapour difference} = \text{Upper water vapour} - \text{Lower water vapour} \tag{1}$$

Channels 8 and 10 represent the upper water vapour and lower water vapour levels respectively. Applying Equation 1 to the channel coefficients we can infer the model's coefficients for WVD, in Figure 11. The reducing coefficient value of the WVD frames implies that cooling of WVD is indicative of DCI. This result was found in the theory in Section 2.1.

4.4.2 Violin plot

The land and sea models were cross-validated to create a coefficient distribution. Non-sparse coefficients are shown as a violin plot in Figure 12. This displays a more general distribution of coefficient values for each channel. The large hyper-parameter value, C, that the sea model was trained with means there are many non-zero, low magnitude coefficients that are plotted. The sparse nature of the land model coefficients means channel trends are clearer to see in Figure 12 for the land domain than the sea domain. Channels 9 and 14 of the land models both show clear negative biases and channels 10 and 15 have positive biases. Implying low values of channels 9 (Midlevel water vapour) and 14 (Longwave window) and high values of 10 (Lower-level water vapour) and 15 ("dirty" longwave window) all contribute to DCI.

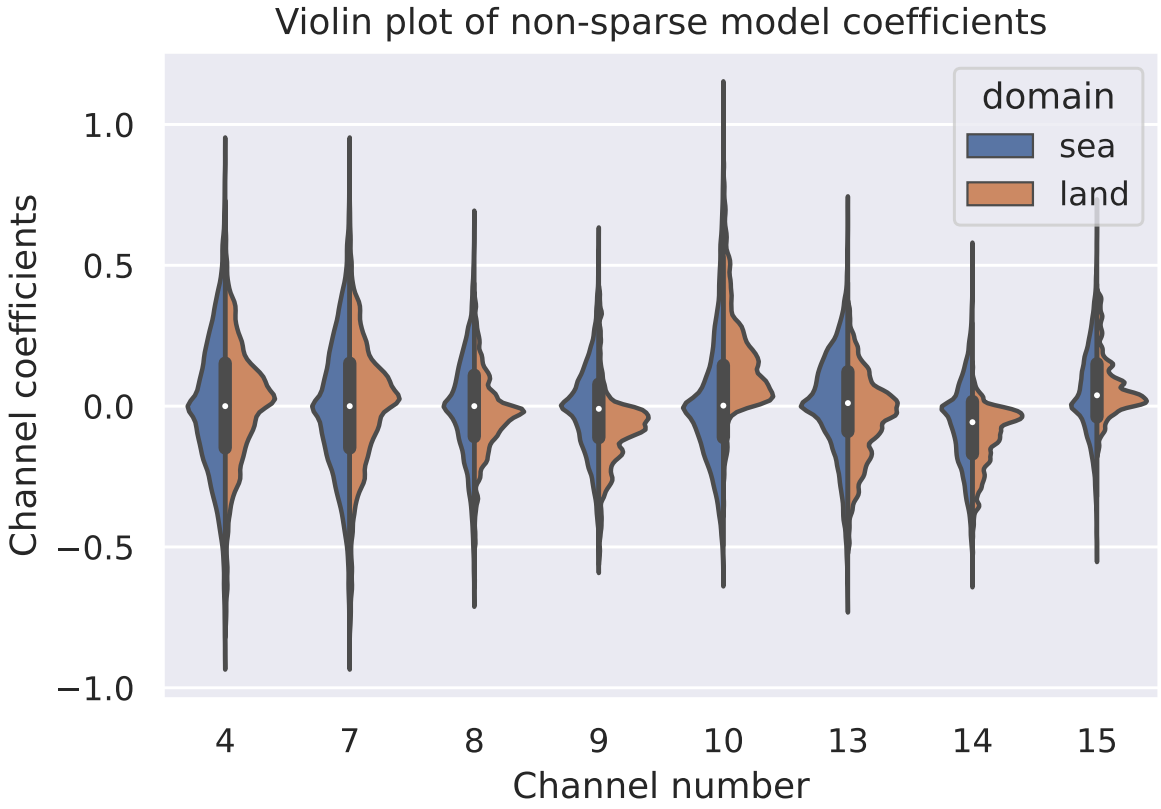

Figure 12: Violin plot of channel coefficients

5 Further work and Conclusion

The linear logistic regression models developed on the data sets extracted from the GOES data successfully labelled both cases of DCI and non-DCI with an accuracy of ∼80%. The models developed for identification over land consistently outperformed those for detection over sea. The 'Cirrus', 'Upper-level water vapour', 'Low-level water vapour' and 'Shortwave' windows were found to be the most influential on labelling deep convective initiation. An estimated influential domain on DCI was set at 10kmx10km over land and 14kmx14km over sea. The leading time that is influential to DCI was infered to be 15 minutes. Analysis of model coefficients found that cooling in the WVD field implies it is a DCI. This is a known mechanism for classifying deep convection. Finding this in the model gives further credibility to model results. Further work could included analysing the performance of the algorithm over different GOES domains, such as central continental United states (CONUS) and south America. A new distinction between land, sea and coastal regions to ensure different convection mechanisms are considered.

References

[1] W. K. Jones, M. W. Christensen, and P. Stier. A semi-lagrangian method for detecting and tracking deep convective clouds in geostationary satellite observations. *Atmospheric Measurement Techniques Discussions*, 2022:1–24, 2022.

[2] William K. Jones. Tobac Flow, 9 2022.

[3] Ulrike Lohmann, Felix Luond, and Fabian Mahrt. *An introduction to clouds*. 2016.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot,

and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[5] Yutaka Sasaki. The truth of the f-measure. *Teach Tutor Mater*, 01 2007.

[6] Timothy J. Schmit, Scott S. Lindstrom, Jordan J. Gerth, and Mathew M. Gunshor. Applications of the 16 spectral bands on the advanced baseline imager (abi). *Journal of Operational Meteorology*, 06(04):33–46, 2018.