Bayesian Unsupervised Machine Learning Approach to Segment Arctic Sea Ice from SMOS

Herbert Christoph[1], Camps Adriano[1], Wellmann Florian[2], and Vall-Llossera Mercedes[1]

[1]Universitat Politècnica de Catalunya (UPC)
[2]RWTH Aachen University

November 16, 2022

Abstract

Microwave radiometry at L-band is sensitive to sea ice thickness (SIT) up to 60 cm. Current methods to infer SIT depend on ice-physical properties and data provided by the ESA's Soil Moisture and Ocean Salinity (SMOS) mission. However, retrieval accuracy is limited due to seasonally and regionally variable surface conditions during the formation and melting of sea ice. In this work, Arctic sea ice is segmented using a Bayesian unsupervised learning algorithm aiming to recognize spatial patterns by harnessing multi-incidence angle brightness temperature observations. The approach considers both statistical characteristics and spatial correlations of the observations. The temporal stability and separability of classes are analyzed to distinguish ambiguous from well-determined regions. Model uncertainty is quantified from class membership probabilities using information entropy. The presented approach opens up a new scope to improve current SIT retrieval algorithms, and can be particularly beneficial to investigate merged satellite products.

Hosted file

`animation_a1.gif` available at https://authorea.com/users/525020/articles/595607-bayesian-unsupervised-machine-learning-approach-to-segment-arctic-sea-ice-from-smos

Bayesian Unsupervised Machine Learning Approach to Segment Arctic Sea Ice from SMOS

Christoph Herbert[1,2], Adriano Camps[1,2], Florian Wellmann[3], and Mercedes Vall-llossera[1,2]

[1]CommSensLab, Universitat Politècnica de Catalunya (UPC) and Institut d'Estudis Espacials de Catalunya (IEEC/CTE-UPC), Barcelona, Spain
[2]Barcelona Expert Center (BEC), Barcelona, Spain
[3]Institute for Computational Geoscience and Reservoir Engineering, RWTH Aachen University, Aachen, Germany

Key Points:

- Retrieval algorithms to infer ice properties, such as sea ice thickness, exhibit high uncertainty due to limited knowledge of complexity
- An Unsupervised learning approach provides a synergistic framework which links data with the aim to recognize and analyze spatial patterns
- Bayesian segmentation of Arctic sea ice from SMOS data reveals stable and separable classes while indicating model uncertainty

Corresponding author: Christoph Herbert, `herbert@tsc.upc.edu`

Abstract

Microwave radiometry at L-band is sensitive to sea ice thickness (SIT) up to $\sim 60\,\mathrm{cm}$. Current methods to infer SIT depend on ice-physical properties and data provided by the ESA's Soil Moisture and Ocean Salinity (SMOS) mission. However, retrieval accuracy is limited due to seasonally and regionally variable surface conditions during the formation and melting of sea ice. In this work, Arctic sea ice is segmented using a Bayesian unsupervised learning algorithm aiming to recognize spatial patterns by harnessing multi-incidence angle brightness temperature observations. The approach considers both statistical characteristics and spatial correlations of the observations. The temporal stability and separability of classes are analyzed to distinguish ambiguous from well-determined regions. Model uncertainty is quantified from class membership probabilities using information entropy. The presented approach opens up a new scope to improve current SIT retrieval algorithms, and can be particularly beneficial to investigate merged satellite products.

Plain Language Summary

Remote sensing techniques are commonly used to provide maps of sea ice thickness (SIT). Methods to obtain these maps are based on the sea ice composition and on the signal measured by satellite. Sea ice Composition is spatially complex and changes during its formation and melting. Currently used data from observations of ESA's Soil Moisture and Ocean Salinity (SMOS) mission depend on several sea ice parameters, which hinders good estimation of almost any specific sea ice parameter. In this work, a new method to combine the information contained in SMOS brightness temperature data is investigated, with the aim to divide the Arctic region into a number of smaller areas – so called classes. Useful information about sea ice is contained in the spatial and statistical distribution of SMOS data, which are collected at different incidence angles. The relationship between the observations and the statistical properties of the obtained classes allow an assessment of its degree of separability and uncertainty. How classes change in time is used to estimate their temporal stability. The presented approach can be used to investigate the link between a variety of spatial datasets to improve current SIT products, and can be applied in many scientific fields.

1 Introduction

The Arctic region shows strong positive feedback to global warming and is very sensitive to climate change. Arctic sea ice has been declining, with the sea ice minimum for September 2020 ending up being the second lowest in the 42-year satellite record (NSIDC, 2020). Sea ice governs heat transfer and influences atmospheric circulation, which is particularly important because low- and mid-latitude's climates are closely related to polar climate (Overland & Wang, 2010; Francis & Vavrus, 2012). Monitoring of both sea ice concentration (SIC), as the fraction of sea-ice cover within an observed cell, and sea ice thickness (SIT) are necessary for a consistent determination of sea ice dynamics. Microwave radiometry is independent of daylight and at lower microwave frequency it is mostly unaffected by atmospheric conditions. The emissivity in the microwave spectrum depends on the dielectric properties of sea ice, which are a function of its physical composition including salinity, density, surface temperature, and surface roughness. In addition, the signal is emitted from a radiating layer which depends on the penetration depth of the sensor. Therefore, the separability of surface properties, such as open water and sea ice including SIT, is - in theory - feasible.

Several algorithms to retrieve SIT and SIC from brightness temperature (T_b) of satellite observations at Arctic scale have been developed, and various products have been deployed (Huntemann et al., 2014; Tian-Kunze et al., 2014; Kaleschke et al., 2016; Ricker et al., 2017; Gupta et al., 2019; Lavergne et al., 2019). ESA's Soil Moisture Ocean Salinity (SMOS) mission (Font et al., 2009; Kerr et al., 2010) provides multi-incidence angle full-polarization T_b maps at L-band (1.4 GHz), which show sensitivity to thin sea ice. However, sea ice is under continuous transformation showing regional and seasonal variability. Physics-based methods to retrieve SIT strongly rely on knowledge of the ice-physical parameters. These parameters are estimated from empirically determined properties of different ice types (e.g. first- or multi-year ice). Thus, models can be subject to over-simplification, and model uncertainty is difficult to estimate, especially at Arctic scales considering an entire year. Validation capability is also limited due to sparsely available, only regionally and seasonally acquired, in-situ and airborne data. SIT retrieval algorithms perform well during Arctic freeze-up (Kaleschke et al., 2016), whereas heterogeneous conditions of sea ice during summer melt and limited spatial resolution of satellite observations make SIT estimation highly ambiguous. Therefore, SIT maps of sufficient quality are only available from mid-October to mid-April.

In this study, a data-based approach is investigated to segment Arctic sea ice, assuming that independent information about its properties are captured in the SMOS multi-incidence angle T_b dataset. The aim is to yield a framework to reveal spatial patterns from differences and similarities in the sensitivity of T_b observations to sea ice properties using an unsupervised learning algorithm. A Bayesian inferential model based on Gaussian Mixture Models (GMM) and Hidden Markov Random Fields (HMRF) considers both the statistical characteristics and the spatial correlations of the observations (Wang et al., 2017). The Arctic region is reduced to a relevant number of spatial classes, while keeping the probabilistic distribution for subsequent cluster analysis and uncertainty quantification. Spatial information is provided in terms of a latent field in physical space and statistical information is indicated by the means and covariances of the obtained classes in the feature space. A direct inference of sea ice properties, particularly at the ocean-ice-boundary, is ambiguous because SMOS observations can be sensitive to both SIC and thin sea ice. Therefore, T_b observation consisting of open water, and low SIC are corrected using SIC maps of the OSI-401-b product, provided by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT). The polarization ratio (PR) between horizontally and vertically polarized values is selected for segmentation to increase the sensitivity to sea ice signatures by reducing the effect of physical surface temperature.

2 Data and Methods

In this study, PR maps at multi-incidence angles are obtained from SMOS T_b observations and OSI-401-b SIC maps, and are used to segment the Arctic ocean into sub-regions based on different sea ice properties. The proposed unsupervised machine learning approach is based on a Bayesian inference framework (Wang et al., 2017). The aim is to indicate patterns in a latent field in physical space according to the most relevant T_b observations. The temporal evolution of these patterns can be analyzed in terms of cluster separability and correlation of the input features to investigate the corresponding sea ice signatures.

2.1 SMOS multi-incidence angle T_b data

ESA's SMOS mission was originally designed to provide global and frequent maps of soil moisture and ocean salinity, but measurements also show sensitivity to different sea ice properties (thin SIT and SIC). The SMOS satellite is equipped with the Microwave Imaging Radiometer with Aperture Synthesis (MIRAS), an interferometric radiometer operating at L-band ($\sim$1.4 GHz) that acquires multi-incidence angle (0 - 60°) full polarization T_b in ascending (6 a.m.) and descending (6 p.m.) sun-synchronous orbit (Corbella et al., 2005). T_b maps are retrieved with a radiometric resolution between 0.8 - 2.2 K, a spatial resolution of $\sim$35 km at centre of field of view, and a revisit time of $\sim$1 - 3 days

(Famiglietti et al., 2008). The retrograde polar orbit (98.42° inclination and 758 km altitude) limits the observations to a maximum latitude of $\sim 84°$, resulting in missing values around the poles ('polar hole'). The input dataset for this study is given by the SMOS Level 1B data product consisting of the Fourier components of T_b in the antenna polarisation reference frame. The high jump discontinuities in T_b between land and sea observations lead to oscillations after image reconstruction at coastal areas (Gibbs phenomenon). These contaminated zones, as well as continental land mass, were removed in the data product. Ascending and descending SMOS observations show only small differences in T_b. Therefore, T_b of both orbits are averaged. A daily multi-angular dataset with 2° sampling is created similar to Gabarró et al. (2016) with T_b provided in horizontal and vertical polarization.

2.2 Input features selection

The study period includes the late summer melt and the first half of the freeze up period from September 1 to December 31, 2016. T_b data are averaged over 5 days to guarantee full coverage of the Arctic ocean. Pixels of T_b images either consist of sea ice with ($0 < SIC \leq 1$), or purely consist of open water (SIT = 0). The sea ice surface represents a grey body, and T_b is the product of the emissivity (ϵ) and the physical temperature (T_{Phys}), which is non-negligible in the lower microwave spectrum and varies depending on the atmospheric conditions among the Arctic. Therefore, input data for segmentation are selected with the objective to correct for SIC and to reduce the effect of spatial and temporal variability of T_{Phys} on T_b. In addition, direct inference of specific sea ice properties, particularly at the ocean-ice-boundary, is ambiguous by the fact that T_b can be sensitive to both SIC and thin SIT.

In a first step, $T_{b_{(SI)}}$ was determined from the observed T_b, SIC, and the freezing point of seawater ($T_{b_{(OW)}}$) (eq. 1).

$$T_b = \alpha\, T_{b_{(SI)}} + (1-\alpha)\, T_{b_{(OW)}} \quad \text{with} \quad \alpha \in [0,1] \quad \text{and} \quad T_b = \epsilon\, T_{phys} \tag{1}$$

Hereby, OSI-401-b SIC maps are provided in a polar stereographic projection grid at 10 km resolution and are regridded and upscaled to SMOS resolution using kd-tree resampling. $T_{b_{(OW)}}$ are determined at different incidence angles and polarizations by evaluating the coldest values obtained for observations with low SIC located at latitudes above 75 °N. SIC is often underestimated with respect to SIT, resulting in an overestimation of T_b, which particularly influences the segmentation of areas covered by thin ice along sea ice edges. Therefore, a SIC threshold of α=0.5 was chosen to provide an open water mask and to exclude observations classified with low SIC, which limits the overestimation error.

In a second step, to account for variations in T_{Phys}, the polarization ratio (PR) is computed as the normalized difference between vertically and horizontally polarized values ($T_{b_{(SI,V)}}$ and $T_{b_{(SI,H)}}$) as follows

$$PR = \frac{T_{b_{(SI,V)}} - T_{b_{(SI,H)}}}{T_{b_{(SI,V)}} + T_{b_{(SI,H)}}} = \frac{\epsilon_{(SI,V)} - \epsilon_{(SI,H)}}{\epsilon_{(SI,V)} + \epsilon_{(SI,H)}}, \tag{2}$$

which reduces to the emissivities of sea ice with the advantage of enhancing the sensitivity to the actual sea ice properties. $T_{b_{(SI,V)}}$ is higher than $T_{b_{(SI,H)}}$ with larger differences for increasing incidence angles. Also, emissivity depends on the optical path length through sea ice, and PR increases for observations at higher incidence angles. PR's obtained for high incidence angles showed sufficient sensitivity range over ice-covered area with values reaching from 0 (thick ice, saturation) to ~ 0.3 (thin ice) and its distribution depends on the observed period. Selecting PR values for high angles increases the content of independent information about sea ice, whereas values at lower angles are more likely to contain redundant information, which may lead segmentation biases. An assessment of the dominant features of SMOS data showed that sufficient angular variability of SMOS T_b can be already obtained using three incidence angles. Therefore, PR maps at 40°, 48° and 56° are used as input features for segmentation.

2.3 Bayesian unsupervised machine learning algorithm

A Bayesian unsupervised machine learning approach Wang et al. (2017) is employed, previously applied to extract patterns of subsurface heterogeneity from geophysical multi-source data (Wang et al., 2019; Herbert et al., 2019). A Gaussian Mixture Model (GMM) is used to fit N data points (image pixels) in an M-dimensional space (M number of features) to find an optimal set of multivariate Gaussian distributions (L classes). The distributions are parametrized by their means $\mu_{\theta,l}$ and covariances $\Sigma_{\theta,l}$ for each cluster l and incidence angle θ. Since features originate from satellite observations, a Hidden Markov Random Field (HMRF) is used to consider the statistical characteristics of data points in feature space as well as their spatial dependencies. A directional smoothing coefficient β accounts for anisotropy conditions with the assumption that neighboring pixels are more likely to belong to the same class. The segmentation results in a latent field x of hidden variables, which indicates the most probable class membership as well as the probability $p(x_i)_l$ of each pixel i to belong to class $l \in L$. The segmentation procedure is described in detail in Wang et al. (2017). The model parameters $(\mu, \mathbf{\Sigma}, \beta)$ as well as the latent field x are obtained through Bayesian optimization in an iterative sampling process using a Markov Chain Monte Carlo (MCMC) approach after an initial Expectation-Maximization step. Prior to segmentation, the number of classes was predefined regarding the distribution of PR values. During late summer melt, only two significant classes are expected, comprising the remaining thick multi-year ice and regions of thinner ice. After sea ice minimum in mid-September, an additional third class is introduced, representing newly formed sea ice during freeze up. This choice is further approved by an a posteriori evaluation of cluster separability.

2.4 Cluster analysis

Results of the Bayesian segmentation are analyzed regarding the obtained patterns in physical space, and the location and orientation of clusters in feature space. The information-theoretic measure of entropy (H) is used to provide model uncertainty. It was initially defined by (Shannon, 1948) in the context of communication and has since been adapted to geosciences (Goodchild et al., 1994; Wellmann & Regenauer-Lieb, 2012). It is used to distinguish well-classified from uncertain regions and is defined by

$$H(x_i) = -\sum_{l=1}^{L} p(x_i)_l \, log(p(x_i)_l), \tag{3}$$

where $p(x_i)_l$ denotes the probability in the physical space of pixel i to belong to class l. H can reach values close to zero (pixel clearly assigned to one class) and $H_{max} = L[1 - log(L)]$ (uniform distribution for L classes).

Sea ice properties, to which SMOS multi-incidence T_b are sensitive to, are dissimilar between classes and show similarities within the same class. Clusters in feature space are investigated regarding their location and orientation by analyzing the model parameters μ and $\mathbf{\Sigma}$. The correlation coefficient ρ quantifies the intra-cluster cohesion and can be used to distinguish between informative and redundant observations (Benesty et al., 2009). It is derived for each cluster from $\mathbf{\Sigma}$ in two-dimensional marginal space between features j and $k \in \{40^\circ, 48^\circ, 56^\circ\}$

$$\mathbf{\Sigma}_l = \begin{bmatrix} \Sigma_{jj} & \Sigma_{jk} \\ \Sigma_{kj} & \Sigma_{kk} \end{bmatrix}_l = \begin{bmatrix} \sigma_j^2 & \sigma_j \, \sigma_k \rho_{jk} \\ \sigma_k \, \sigma_j \rho_{kj} & \sigma_k^2 \end{bmatrix}_l, \tag{4}$$

where $\sigma_{j,k}$ correspond to the standard deviations with respect to feature j, k and $\rho_{jk} = \rho_{kj}$ denote the correlation coefficients between two features, given by

$$\rho_{jk} = \frac{\Sigma_{jk}}{\sigma_j \sigma_k} = \frac{\Sigma_{jk}}{\Sigma_{jj}^{1/2} \Sigma_{kk}^{1/2}}, \quad -1 \leq \rho_{jk} \leq 1. \tag{5}$$

The Geometric Separability Index (GSI) (Thornton, 1998) is a distance-based measure to analyze inter-cluster separability and is widely used for cluster interpretation (Greene, 2001; Mthembu & Marwala, 2008). GSI compares all N data points with their nearest neighbor regarding their class membership and is defined by

$$\mathrm{GSI}(f) = \sum_{i=1}^{N} \frac{(f(x_i) + f(x_i') + 1) \bmod 2}{N} \quad \text{with} \quad f(x_i) = \begin{cases} 1, & \text{if } x_i' = x_i \\ 0, & \text{if } x_i' \neq x_i \end{cases}, \tag{6}$$

where f is a binary target function, and x_i' is the nearest neighbor of x_i in the feature space of pixel i. $\mathrm{GSI} \in [0.5, 1]$ and for values reaching its lower or upper limit, clusters are completely entangled or ideally separable, respectively. In this study, both global and cluster-specific separability are estimated. Global separability is computed based on Euclidean distance for all data points, and cluster-specific separability is obtained based on Mahalanobis distances $(x_i - \mu)\, \boldsymbol{\Sigma}^{-1}\, (x_j - \mu)^T$, considering the data points and covariances of the specific cluster (Mahalanobis, 1936). GSI is investigated along the study period to evaluate the dynamics of the underlying sea ice properties and the stability of the segmentation.

3 Results

Arctic sea ice is segmented independently for 5-day intervals into classes during the periods of late summer melt and early freeze up from September 1 to December 31, 2016. The latent field in physical space and the corresponding multivariate Gaussian distributions of data points in feature space are presented as an example for the segmentation step interval between October 24 - 28, 2016 (sections 3.2 and 3.1). The temporal evolution of model parameters (cluster means and variation) is evaluated in section 3.3. Class membership and separability are assessed in section 3.4 to indicate cluster stability and performance of the algorithm.

3.1 Latent field of classes in physical space

The Figures 1a and 1b show the resulting latent field and the model uncertainty quantified by information entropy, respectively. The latent field indicates spatial patterns, which are acquired from the final iteration of the segmentation by assigning the class with highest probability to every pixel. Pixels with the probability to belong to two or more clusters have larger entropy and reflect therefore uncertain pixels. These pixels comprise regions at the boundary between classes and pixels, which are generally difficult to assign to any cluster. In the latter case, these pixels may point out sub-regions with different sea ice properties (anomalies), which are characterized with high model uncertainty.

The segmented spatial patterns are compared to those of the SMOS L3 Sea Ice Thickness product, provided by the Alfred Wegener Institute (AWI) for Polar and Marine Research (Tian-Kunze et al., 2014). SIT means were computed according to the indicated spatial classes in each segmentation step, and averaged values are determined during freeze up from October 15 to December 31, 2016. The three classes can associated to different ice thickness (in meters) of 1.24 ± 0.10, 0.54 ± 0.24 and 0.13 ± 0.07, respectively. The classes are labeled as ($0 \,\widehat{=}$ thick ice up to sensor saturation), ($1 \,\widehat{=}$ transition zone with higher thickness variability, containing various ice types), and ($2 \,\widehat{=}$ newly-formed thin ice).

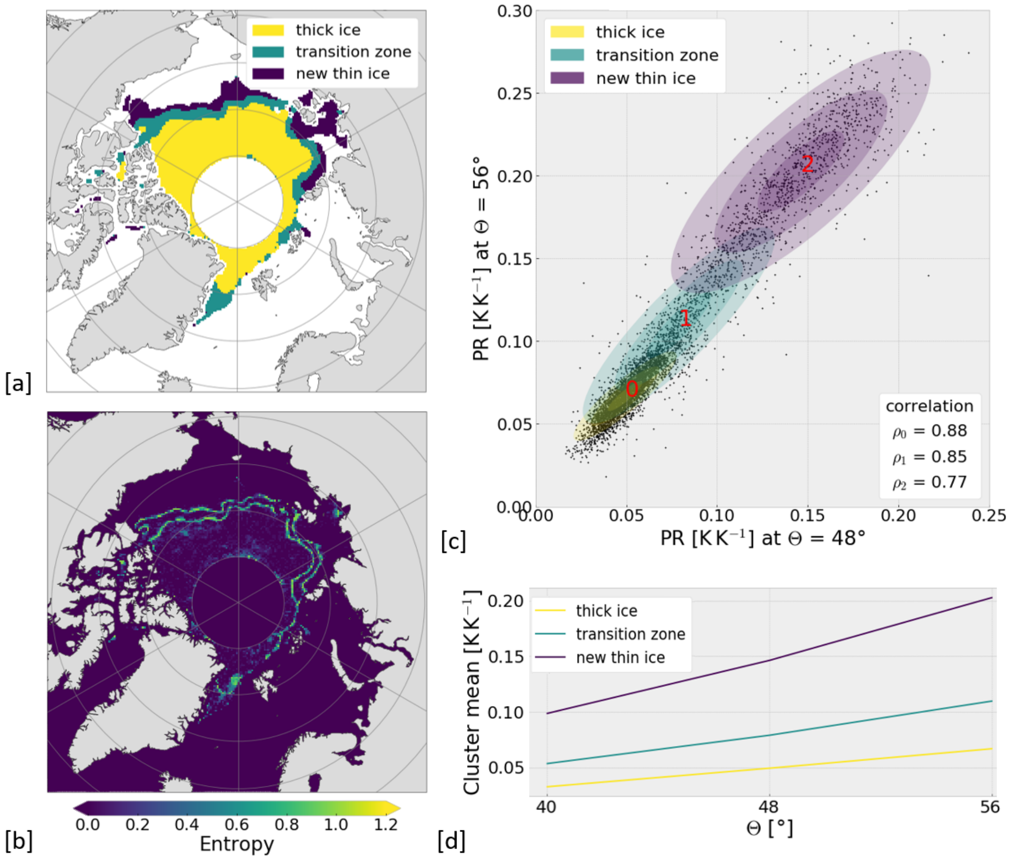

Figure 1. Segmentation result for observations between October 24 and October 28, 2016. [**a**] Latent field result for three classes. [**b**] Model uncertainty represented by information entropy based on label probabilities. [**c**] PR in marginal feature space between the incidence angles 48 and 56 $^\circ$ and correlation for each cluster. [**d**] Variation of PR cluster means with incidence angle.

3.2 Clusters in feature space

Figure 1d illustrates the PR cluster means for different incidence angles. Higher values are obtained for higher incidence angles, characterized by different slopes within the same class, showing that the set of selected input features provides independent information about the sea ice surface. The multivariate Gaussian distributions with the corresponding clusters in marginal features space between the incidence angles 48 and 56 $^\circ$ are illustrated in figure 1c. The correlation between the input features is generally higher for thick ice resulting in a well-determined cluster with higher intra-cluster cohesion. In contrast, newly-formed thinner ice shows less correlation between input features. This enables to discriminate classes of similar surface characteristics, to which multi-incidence angle observations show a different signature. However, sea ice is a complex medium and sea ice growth can occur under rougher or calmer ocean conditions, causing newly formed ice to be heterogeneous. These differences in the origin of sea ice formation might be captured in the input features, indicated by a broader distribution in marginal features space. On the contrary, the structures of multi-year thick ice appear more homogeneous.

3.3 Temporal evolution of clusters

Figure 2 shows the temporal evolution of cluster means and standard deviations (StDev) in marginal feature space for $\theta = 56\,^{\circ}$, and the distribution of PR and the corresponding class membership at three particular dates. The late summer melt comprises two significant classes until annual sea ice extent reaches its minimum (September 6, 2016). The evolution of cluster means is compared to the mean Arctic temperature, which is computed from daily 2 m temperature ERA5 reanalysis data for latitudes above 75 °N and downloaded from the European Centre for Medium-Range Weather Forecasts (ECMWF) (C3S, 2017). Once Arctic temperatures drop long enough below the freezing point of saline sea water ($\sim$-1.8 $°C$) to allow sufficient heat transfer towards the atmosphere, new sea ice starts to form. Hence, a third class can be determined, which is represented by a significant number of PR values above 0.15. Cluster mean of thick ice is widely stable over the entire study period. Two phenomena can be observed regarding new thin ice. Firstly, its cluster mean decreases and gradually closes up with the transition zone. Secondly, an overlap between clusters can be observed in relation to strong positive temperature anomalies in the Arctic. This can be due to class imbalance arising from a decreasing amount of newly-formed ice, comparing to the total sea ice extent. Also, as the sea ice edge reaches lower latitudes during freeze up, which are characterized by different climate conditions, a decrease in PR values can be observed although a significant amount of sea ice is still being formed.

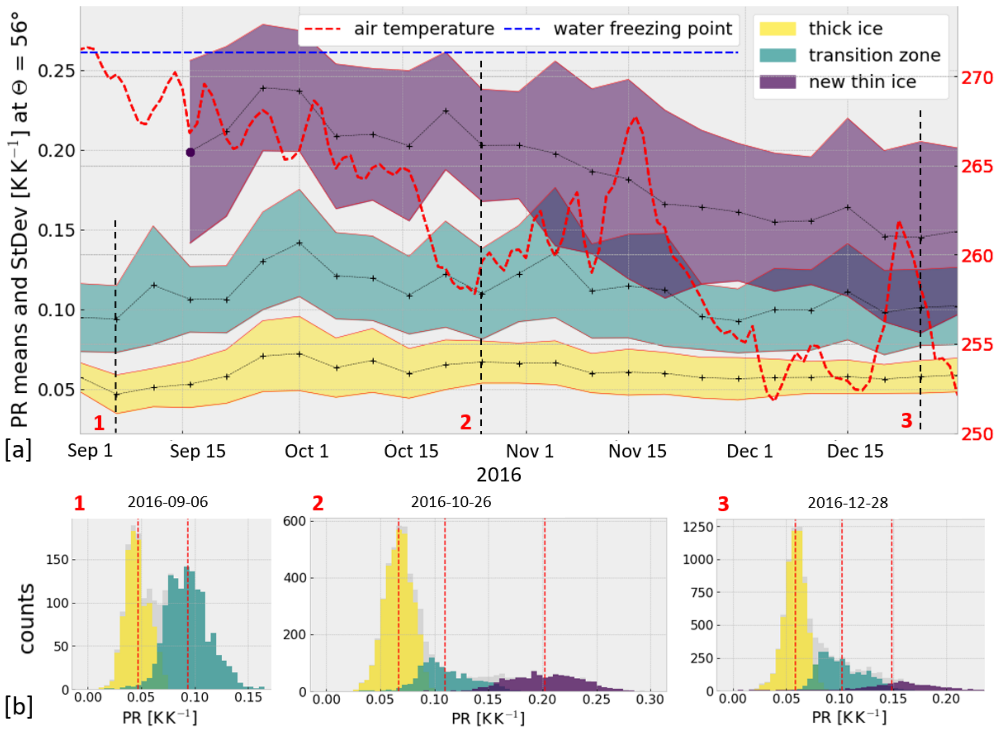

Figure 2. Temporal evolution of clusters. [**a**] Temporal evolution of cluster means and standard deviations at 56 ° incidence angle and mean Arctic temperature for latitudes > 75 °N. [**b**] PR distribution with respect to class membership at three particular dates (September 4-8, October 24-28 and December 24-28, 2016).

Figure 3a shows the evolution of the number of pixels per class membership for filtered sea ice (SIC > 0.5) in comparison to the total sea ice extent (SIE). SIE comprises sea ice cover for SIC > 0.15 and daily data was downloaded from the data archive of the National Snow and Ice Data Center (NSIDC, 2020). Deviations and offsets between SIE and the total pixel counts are due to missing values within the 'polar hole' and contaminated zones at the sea-land boundary. The increase of the total number of pixels is equivalent to a monthly growth rate in SIE of about $2.5 \times 10^6\,\mathrm{km}^2$. The number of pixels consisting of newly-formed ice is broadly stable, whereas the number pixels classified as transition zone are slightly increasing during freeze up. As sea ice grows, thick sea ice becomes more abundant, leading to a log-normal-shaped PR distribution with increasing expected value (Figure 2b,3). Although thin ice becomes less representative in the data during freeze up, the algorithm is still capable of separating three classes as long as sea ice formation continues.

3.4 Separability of clusters

Global and cluster-specific separability are shown in figure 3b. The solid lines show the GSI for a choice of two classes in late summer melt and three classes during freeze up. High global separability is achieved along the entire study period with values around 0.9. The cluster-specific GSI indicates separable classes with mean values of 0.95, 0.83 and 0.83 for thick ice, transition zone and new sea ice, respectively. Along the freeze up, new thin ice starts to overlap with the transition zone and a threshold of minimum GSI needs to be defined to specify the appropriate number of classes for each segmentation step. For comparison, GSI is shown for the end of the summer melt period for a segmentation with three classes (dashed lines). In this case, classes highly overlap and the choice of two initial clusters from the beginning of the study period leads to higher separability.

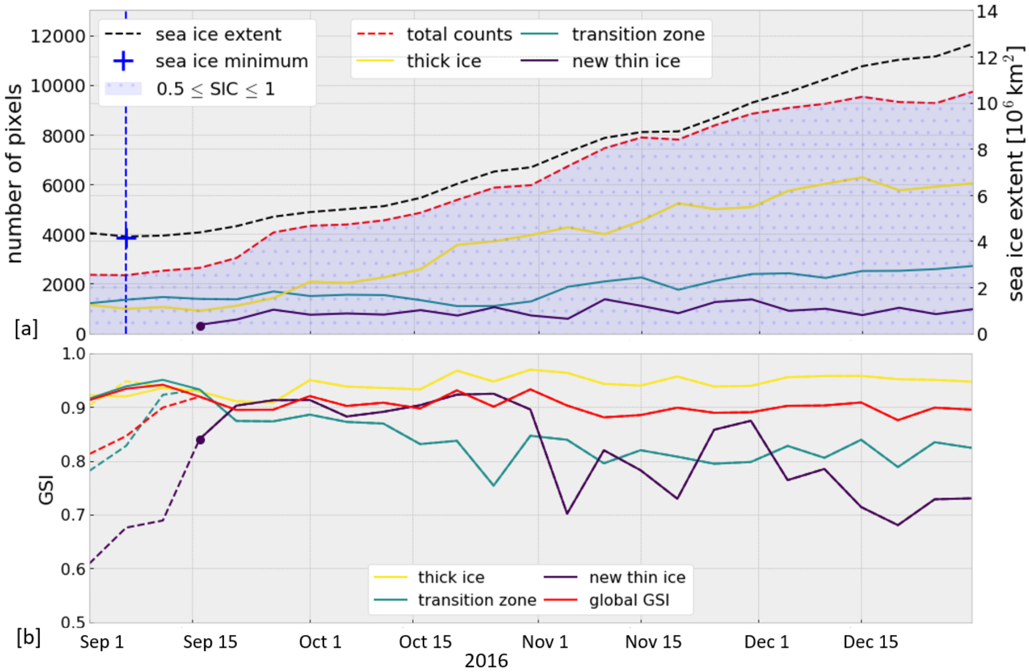

Figure 3. [**a**] Temporal evolution of class membership and sea ice extent, with indicated sea ice minimum and SIC. [**b**] Global and cluster-specific GSI along the observation period, determined from nearest-neighbor evaluation using Euclidean and Mahanalobis distances, respectively.

4 Discussion

A novel approach is evaluated to obtain sea ice maps from SMOS observations using Bayesian segmentation. The estimation of a constant number of stable and separable classes revealed periods, when T_b observations show similar sea ice signatures. The information content obtained by linking T_b data at multiple incidence angles and polarizations is reduced to a number of most significant classes, with good inter-cluster separability. The corresponding spatial patterns, which are indicated in the latent field result, can be used to extract the heterogeneity of the underlying sea ice properties.

Information entropy points out both uncertain zones between segmented classes and anomalies which can form sub-classes. As an example, ponded sea ice during summer melt has different surface characteristics, which may result in a further discriminable class only during that particular period. Since cluster means represent the most significant observations at every segmentation step, their temporal evolution can be used to define dynamic tie points. These tie points can be analyzed to investigate how sensitive input features respond to changes in sea ice signatures.

The implemented method serves as a framework to integrate multi-source datasets and is capable of recognizing patterns by considering the statistical characteristics and spatial correlations. The relationship of satellite observations at multiple frequencies can be used to select an appropriate set of input features and to enhance the sensitivity to ice-physical parameters, such as SIT. A combination of the presented data-driven segmentation approach with a physics-based inference model build upon the estimated distribution of classes may increase the retrieval accuracy of existing large-scale sea ice products.

5 Conclusion

In this work, Arctic sea ice is classified using a Bayesian unsupervised learning approach by making full use of the information about sea ice properties contained in the PR of SMOS multi-incidence angle T_b data. Sea ice properties are considered anisotropic as well as regionally and seasonally variable among the Arctic and T_b cannot be assumed to be sensitive to similar properties over an entire year. Therefore, both statistical characteristics of observations are evaluated and the segmentation is carried out by means of a discretized number of spatially regularized classes. The number of classes was determined a priori from the PR distribution and was verified a posteriori using GSI. Model uncertainty was determined using information entropy and enabled to distinguish well-determined from uncertain regions. High global separability was achieved considering two classes during late summer melt and three classes during freeze up, respectively. A comparison with existing SMOS-SIT maps indicated that classes can be attributed to SIT ranges. During late summer melt, two classes could be attributed to remaining thick ice and a transition zone, showing differences in the correlations of the input features. With the beginning of the formation of new thin ice during freeze up, an additional class could be discriminated based on the occurrence of higher PR values. However, the decrease in relative abundance of newly formed ice to the total sea ice during freeze up resulted thin sea ice to be less significant and led to higher overlap between classes. The underlying sea ice properties and the corresponding variation in PR have to be better understood to draw conclusions of the obtained classes, considering an entire annual cycle of Arctic sea ice formation and melting.

Acknowledgments

There are no perceived conflicts of interest for the lead author or coauthors. The lead author received the support of a fellowship from “la Caixa” Foundation (ID 100010434). The fellowship code is LCF/BQ/DI18/11660050. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 713673. It was also funded through the award "Unidad de Excelencia María de Maeztu" MDM-2016-0600, by the Spanish Ministry of Science and Innovation through the project "L-band" ESP2017-89463-C3-2-R, and the project "Sensing with Pioneering Opportunistic Techniques (SPOT)" RTI2018-099008-B-C21/AEI/10.13039/501100011033. The authors would also like to thank Carolina Gabarro from the Barcelona Expert Center for providing the processed SMOS L1C data. The production of the SMOS-SIT data was funded by the ESA project SMOS & CryoSat-2 Sea Ice Data Product Processing and Dissemination Service, and data from October 15 to December 31, 2016 were obtained from AWI (`https://smos-diss.eo.esa.int/socat/L3_SIT_Open`).

References

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1–4). Springer.

C3S. (2017). *ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.* `https://cds.climate.copernicus.eu/cdsapp#!/home`. Copernicus Climate Change Service Climate Data Store (CDS). (Accessed: 2020-01-10)

Corbella, I., Torres, F., Camps, A., Colliander, A., Martín-Neira, M., Ribó, S., ... Vall-llossera, M. (2005). Miras end-to-end calibration: Application to smos l1 processor. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(5), 1126–1134.

Famiglietti, J. S., Ryu, D., Berg, A. A., Rodell, M., & Jackson, T. J. (2008). Field observations of soil moisture variability across scales. *Water Resources Research*, *44*(1).

Font, J., Camps, A., Borges, A., Martín-Neira, M., Boutin, J., Reul, N., ... Mecklenburg, S. (2009). Smos: The challenging sea surface salinity measurement from space. *Proceedings of the IEEE*, *98*(5), 649–665.

Francis, J. A., & Vavrus, S. J. (2012). Evidence linking arctic amplification to extreme weather in mid-latitudes. *Geophysical research letters*, *39*(6).

Gabarró, C., Pla Resina, J., Turiel, A., Portabella, M., Martínez, J., Olmedo, E., & González, V. (2016). Arctic sea ice concentration estimation with smos data.

Goodchild, M., Chih-Chang, L., & Leung, Y. (1994). Visualizing fuzzy maps. *Visualization in geographical information systems*, 158–167.

Greene, J. (2001). Feature subset selection using thornton's separability index and its applicability to a number of sparse proximity-based classifiers. In *Proceedings of annual symposium of the pattern recognition association of south africa.*

Gupta, M., Gabarro, C., Turiel, A., Portabella, M., & Martinez, J. (2019). On the retrieval of sea-ice thickness using smos polarization differences. *Journal of Glaciology*, *65*(251), 481–493.

Herbert, C., Wellmann, F., Wang, H., & Hebel, C. v. (2019). Extracting heterogeneity of subsoil from geophysical measurements using unsupervised learning algorithms. In *Geophysical research abstracts* (Vol. 21).

Huntemann, M., Heygster, G., Kaleschke, L., Krumpen, T., Mäkynen, M., & Drusch, M. (2014). Empirical sea ice thickness retrieval during the freeze up period from smos high incident angle observations. *The Cryosphere*, *8*(2), 439–451.

Kaleschke, L., Tian-Kunze, X., Maaß, N., Beitsch, A., Wernecke, A., Miernecki, M., ... others (2016). Smos sea ice product: Operational application and validation in the barents sea marginal ice zone. *Remote Sensing of Environment*, *180*, 264–273.

Kerr, Y. H., Waldteufel, P., Wigneron, J.-P., Delwart, S., Cabot, F., Boutin, J., ...

others (2010). The smos mission: New tool for monitoring key elements ofthe global water cycle. *Proceedings of the IEEE*, *98*(5), 666–687.

Lavergne, T., Sørensen, A. M., Kern, S., Tonboe, R., Notz, D., Aaboe, S., . . . others (2019). Version 2 of the eumetsat osi saf and esa cci sea-ice concentration climate data records. *Cryosphere*, *13*(1), 49–78.

Mahalanobis, P. C. (1936). On the generalized distance in statistics..

Mthembu, L., & Marwala, T. (2008). A note on the separability index. *arXiv preprint arXiv:0812.1107*.

NSIDC. (2020). *Arctic sea ice at minimum extent for 2020 (September 2020).* `https://nsidc.org/news/newsroom/arctic-sea-ice-minimum-extent-2020`. Boulder, Colorado USA. NASA National Snow and Ice Data Center. (Accessed: 2020-10-08)

Overland, J. E., & Wang, M. (2010). Large-scale atmospheric circulation changes are associated with the recent loss of arctic sea ice. *Tellus A*, *62*(1), 1–9.

Ricker, R., Hendricks, S., Kaleschke, L., Tian-Kunze, X., King, J., & Haas, C. (2017). A weekly arctic sea-ice thickness data record from merged cryosat-2 and smos satellite data. *The Cryosphere, 11, 1607-1623*. `https://doi.org/10.5194/tc-11-1607-2017`.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Thornton, C. (1998). Separability is a learner's best friend. In *4th neural computation and psychology workshop, london, 9–11 april 1997* (pp. 40–46).

Tian-Kunze, X., Kaleschke, L., Maaß, N., Mäkynen, M., Serra, N., Drusch, M., & Krumpen, T. (2014). Smos-derived thin sea ice thickness: algorithm baseline, product specifications and initial verification. *The Cryosphere*, *8*, 997–1018.

Wang, H., Wellmann, F., Zhang, T., Schaaf, A., Kanig, R. M., Verweij, E., . . . van der Kruk, J. (2019). Pattern extraction of topsoil and subsoil heterogeneity and soil-crop interaction using unsupervised bayesian machine learning: An application to satellite-derived ndvi time series and electromagnetic induction measurements. *Journal of Geophysical Research: Biogeosciences*, *124*(6), 1524–1544.

Wang, H., Wellmann, J. F., Li, Z., Wang, X., & Liang, R. Y. (2017). A segmentation approach for stochastic geological modeling using hidden markov random fields. *Mathematical Geosciences*, *49*(2), 145–177.

Wellmann, J. F., & Regenauer-Lieb, K. (2012). Uncertainties have a meaning: Information entropy as a quality measure for 3-d geological models. *Tectonophysics*, *526*, 207–216.

Supporting Information for "Bayesian Unsupervised Machine Learning Approach to Segment Arctic Sea Ice from SMOS"

Christoph Herbert[1,2], Adriano Camps[1,2], Florian Wellmann[3], and Mercedes Vall-llossera[1,2]

[1]CommSensLab, Universitat Politècnica de Catalunya (UPC) and Institut d'Estudis Espacials de Catalunya (IEEC/CTE-UPC), Barcelona, Spain

[2]Barcelona Expert Center (BEC), Barcelona, Spain

[3]Institute for Computational Geoscience and Reservoir Engineering, RWTH Aachen University, Aachen, Germany

Contents of this file

1. Figures F1 to F2
2. Table T1

Additional Supporting Information (Files uploaded separately)

1. Animation A1

Corresponding author: Christoph Herbert, Department of Signal Theory and Communications (TSC), Technical University of Catalonia, Building D3, Jordi Girona 1-3, 08034 Barcelona, Spain. (herbert@tsc.upc.edu)

Introduction

The choice of the appropriate number of classes is justified from the distribution of input features and the class separability. Figure F1 shows the marginal distribution of the polarization ratio (PR) at an incidence angle of 56° at 4 particular dates including the end of summer melt and the early freeze-up. PR distributions are shown for segmentation with 2 and 3 classes, respectively. During late summer melt until sea ice minimum (September 10, 2016), the choice of two classes was expected from the shape of the PR distribution. With the beginning of the freeze-up period, higher PR values become more frequent and an additional class is expected. Class separability is indicated by the Geometric Separability Index (GSI) and was obtained subsequent to segmentation. From the segmentation step at September 16, 2016 onwards, segmentation with a choice of 3 classes results in higher separability.

The classes were labeled according to the sea ice thickness estimates of the available SMOS-SIT product (Tian-Kunze et al., 2014). Figure F2 visualizes the latent field result in comparison to SMOS-SIT maps at the segmentation step intervals October 19-23, November 8-12, and December 23-27, 2016. Class 0 predominately contains consolidated thick ice beyond the sensitivity range of L-band $>\sim 0.6\,\mathrm{m}$ (sensor saturation), class 1 refers to a transition zone of multiple thickness and types, and class 2 can be attributed to newly-formed thin ice.

Table T1 summarizes the obtained class mean values and standard deviations, averaged over the freeze-up period from October 15 to December 31, 2016. At each segmentation step interval, SIT mean values for each class are calculated according to the spatial patterns of the latent field result. The obtained values at each segmentation step are then averaged over the freeze-up. The classes 0 and 2 show less variation and form stable clusters along the entire period, whereas class 1 contains higher variation. All three classes show sufficient separability along the entire period.

Caption Animation A1 Animation of the spatial patterns of the latent field result in physical space at each segmentation step interval from September 1 to December 31, 2016, including the distribution of the PR at 56° incidence angle.

References

Tian-Kunze, X., Kaleschke, L., Maaß, N., Mäkynen, M., Serra, N., Drusch, M., & Krumpen, T. (2014). Smos-derived thin sea ice thickness: algorithm baseline, product specifications and initial verification. *The Cryosphere*, *8*, 997–1018.

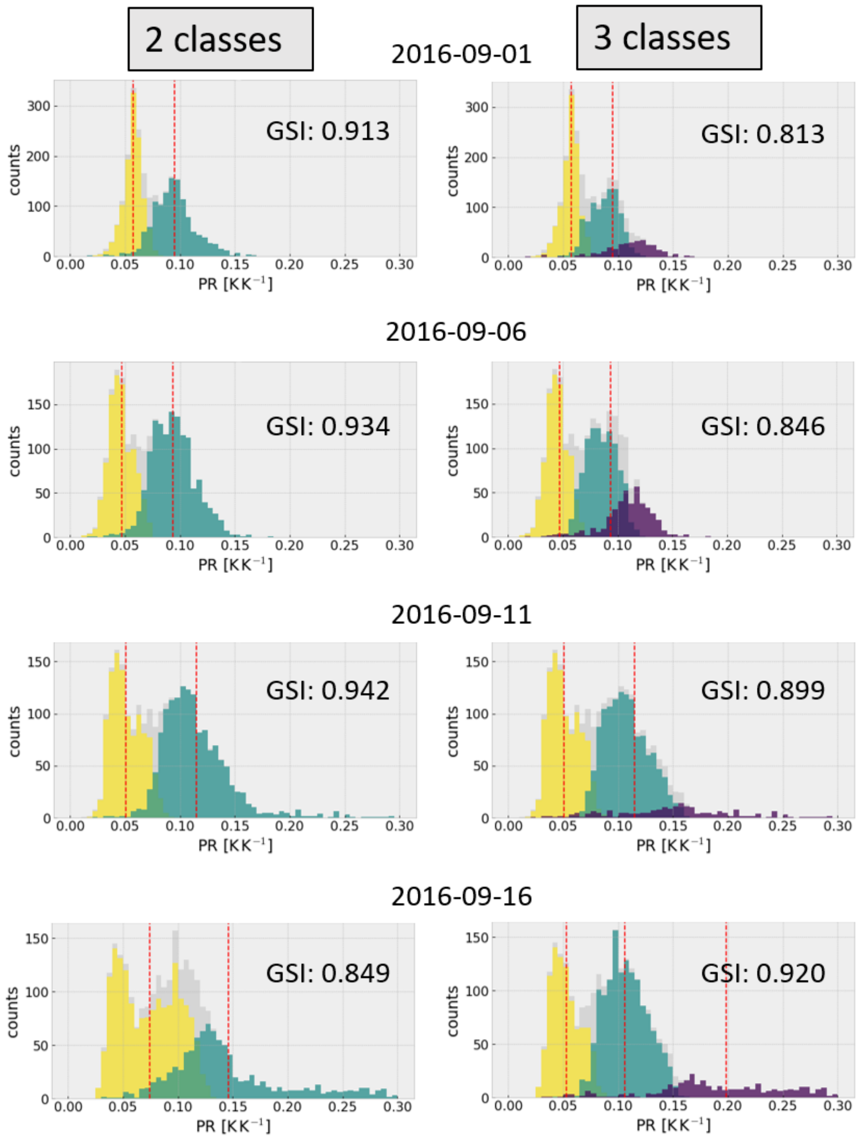

Figure F1. Distribution of PR values at 56° incidence angle for late summer melt and early freeze up from September 1 to September 16, 2016, including the indicated class membership and global separability (GSI), obtained for segmentation with 2 classes (left-hand side) and 3 classes (right-hand side), respectively.

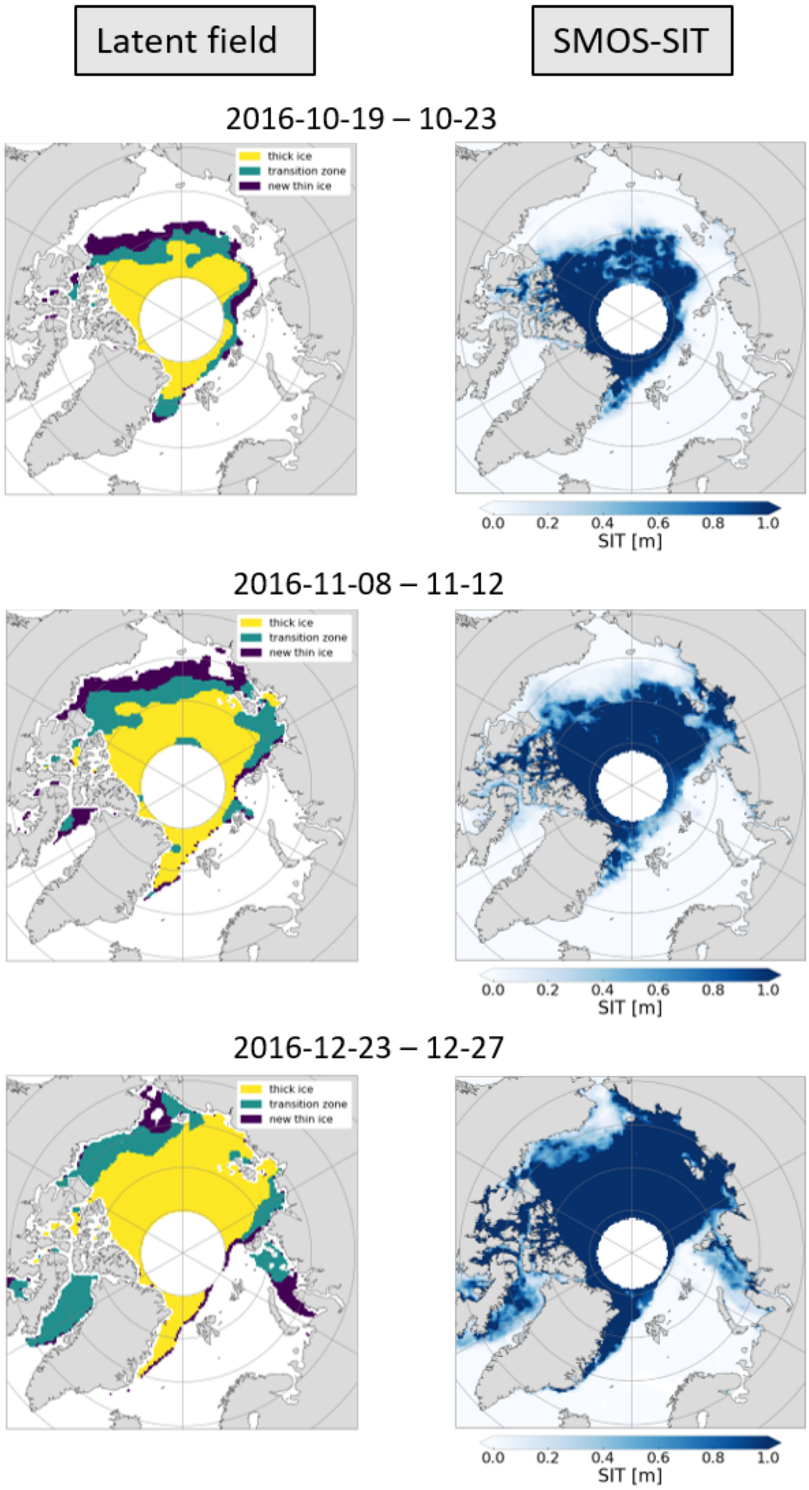

Figure F2. Comparison of the obtained latent field result with SIT maps of the SMOS-SIT product, averaged over the corresponding segmentation period (5-day interval).

Table T1. Summary of the temporal evolution of classes, evaluated within the freeze-up period from October 15 to December 31, 2016. Comparison of PR cluster mean values and standard deviations (StDev) at 56° incidence angle, including global separability (GSI), with the SMOS-SIT product.

Class	Label	PR mean	PR StDev	GSI	SMOS-SIT [m]
0	Thick ice	0.061 ± 0.005	0.014 ± 0.004	0.95 ± 0.02	1.24 ± 0.010
1	Transition zone	0.112 ± 0.012	0.028 ± 0.006	0.83 ± 0.04	0.54 ± 0.24
2	New thin ice	0.187 ± 0.03	0.048 ± 0.009	0.83 ± 0.08	0.13 ± 0.07