Compact models for adaptive sampling in marine robotics

Trygve Olav Fossum¹,², John Ryan³, Tapan Mukerji⁴, Jo Eidsvik⁵, Thom Maughan³, Martin Ludvigsen¹,²,⁶ and Kanna Rajan²,⁷,⁸

Abstract
Finding high-value locations for in situ data collection is of substantial importance in ocean science, where diverse biophysical processes interact to create dynamically evolving phenomena. These cover a variable spatial extent, and are sparse and difficult to predict. Autonomous robotic platforms can sustain themselves in harsh conditions with persistent presence, but require deployment at the correct place and time. To that end, we consider the use of remote sensing data for building compact models that can improve skill in predicting sub-mesoscale features and inform onboard sampling. The model enables prediction of regional patterns based on sparse in situ data, a capability that is essential in regions where use of satellite remote sensing in real time is often limited by cloud cover. Our model is based on classification of sea-surface temperature (SST) images, but the technique is general across any remotely sensed parameter. Images having similar magnitude and spatial patterns are grouped into a compact set of conditional means representing the dominant states. The classification is unsupervised and uses a combination of dictionary learning and hierarchical clustering. The method is demonstrated using SST images from Monterey Bay, California. The consistency of the classification result is verified and compared with oceanographic forcing using historical wind measurements. The established model is then shown to work in a real application using measurements from an autonomous surface vehicle (ASV), together with forecast and sampling strategies. Finally an analysis of the model prediction error is presented and compared across different paths and survey duration.

Keywords
Machine learning, sampling, ocean modeling, marine robotics

1. Introduction
Effective and informative sampling of the ocean requires data gathering strategies that can resolve the spatial and temporal variations of phenomena. This is a formidable challenge owing to the dynamic and unstructured nature of the ocean, with spatio-temporal scales spanning many orders of magnitude, making it unrealistic to observe the dynamics in detail. In addition, coastal waters are often heterogeneous in nature owing to interactions between bathymetry, river discharge, oceanic circulation, as well as endogenous processes (e.g., biology). Methodologically this drives a requirement for using compact spatial models to inform more effective sampling strategies, capable of running on robotic platforms, utilizing prior and current in situ observations. There are numerous ways to build spatial models. The essential goal is to exploit the underlying spatial correlation structures and try to reconstruct the environment, so that future sensing locations can be determined accordingly.

Earth observing satellites offer the possibility to observe a large spatial extent for a range of ocean parameters, such as sea-surface temperature (SST), sea-surface height (SSH), salinity, and ocean color. From such data, a number

¹Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
²Centre for Autonomous Marine Operations and Systems (AMOS), Trondheim, Norway
³Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA
⁴Stanford University, Department of Energy Resources Engineering and Department of Geophysics by courtesy, Stanford, CA, USA
⁵Department of Mathematical Sciences, NTNU, Trondheim, Norway
⁶University Centre in Svalbard (UNIS), Longyearbyen, Norway
⁷Department of Engineering Cybernetics, NTNU, Trondheim, Norway
⁸Underwater Systems and Technology Laboratory, Faculty of Engineering, University of Porto, Portugal

Corresponding author:
Trygve Olav Fossum, Department of Marine Technology, Norwegian University of Science and Technology, Trondheim, Norway.
Email: trygve.o.fossum@ntnu.no
of ocean processes can be discerned and characterized, including algal blooms, fronts, eddies, internal waves, and numerous water quality parameters (Johannessen et al., 2000). Consequently, there is enormous potential for using such information to perform automated analysis of the spatial patterns in the ocean. However, a major challenge using modern machine learning techniques is the reliance on labeled datasets (Gonalves et al., 2008). In remote sensing applications, this challenge is further exacerbated in having to work with a limited number of training samples (Mountrakis et al., 2011), especially when concentrating on a specific area of scientific interest. It is therefore valuable to use unsupervised methods that provide the ability to learn the inherent structure of the data without using explicitly provided labels.

Given the challenge with undersampling in oceanography (Munk, 2002) and the limited availability of accurate real-time information, it is essential to call for capabilities that can estimate spatial and temporal variations in the ocean environment on the fly. Building compact spatial models using prior data sources such as remote sensing and ocean model output is therefore one such possibility. We present a method for building such a model using remotely sensed SST data, as well as examples of how this model can be used in a robotic sensing framework. Remote sensing images are chosen specifically due to their synoptic properties providing repeated large-scale surface observations with reasonably high resolution. The model aims at predicting the current state of the environment using a superposition of states, referred to as classes or scenarios, where each state is established from a similar set of SST images that represent recurring states of oceanographic conditions. The environment is, therefore, assumed to be equal to such a superposition and is found by evaluating the likelihood of these states against observations. A conceptual view of this approach is presented in Figure 1, showing how the model can inform sampling in a sense-plan-act control structure.

It is important that this type of data reduction (unsupervised clustering) cover the common dominant spatial patterns seen in the images, i.e., in the form of several distinct classes of conditional means, that is comparable with the variation in the underlying environment. Using a combination of dictionary learning (Aharon et al., 2006), sparse coding (Mairal et al., 2009), and (agglomerative) hierarchical clustering (Everitt et al., 2011) we propose an unsupervised classifier that can group images with similar oceanographic characteristics, using spatial patterns and the magnitude of temperature from SST images. The idea is not only to automate this process, but also to provide new insight into the underlying processes themselves. Subsequently, the classified SST images are used to distill a compact model of the dominant features by computing conditional means within these classes. One potential drawback to this approach is the lack of uniformity in data acquisition owing to cloud cover and lack of satellite coverage. These factors limit the ability to do repeated and systematic observations of a region of interest, which in the worst case can impede the construction of such spatial models. As a mitigation, data from a 3-year period in spring is used (see Figure 2), such that we obtain a relatively continuous data coverage for a season. Combining this approach with other synoptic sources of data such as numerical ocean models can also be investigated and is discussed further in Section 7.

The applicability of such compact models is an effective characterization of key ocean states, especially when SST data from satellites is not available. In this work, we show an example of this, determining regional conditions for a given day, using data gathered from a WaveGlider ASV (shown in the corner of Figure 3b). We also examine the model prediction error, by computing the expected misclassification rates for different sampling design strategies.

The structure of this article is as follows. In Section 2 we introduce related work, along with some of the challenges that underpin the motivation for this article. Section 3 presents some background on both the data sources and methods used. Section 4 presents the proposed classification methodology, with the results shown in Section 5. Section 6 presents model implementation and usage towards adaptive sampling. We conclude in Section 7 with a summary, followed by a discussion of future work.
2. Related work

Much of the work on automated analysis of remote sensing data is focused on thematic mapping for terrestrial applications, with numerous applications such as in agriculture (Mulla, 2013), mapping of urban environments (Saritha and Kumar, 2017), and vegetation (Xie et al., 2008), and is usually geared towards change detection (Walter, 2004). Typically, the general objective is to categorize regions of an image into one of various land cover classes or themes. Dictionary-based classification has been explored for terrestrial hyper-spectral thematic mapping in Chen et al. (2011). Similar approaches such as sparse reconstruction-based classification have been applied to high-resolution images of the sea floor, taken with synthetic aperture sonar (McKay et al., 2016). Despite an abundance of approaches, Wilkinson (2005) demonstrated through evaluation of 15 years of remote sensing research related to classification, that no technique displays any significant advantage over another, showing no trend in improvement of the classification results, including results from artificial neural networks. Limited contexts, increasing complexity of datasets, lack of embodiment in best practices, the difference between low-level features and high-level user requirements, and imperfect human processes (such as labeling and ground truthing) are some explanations proposed for this lack of progress.

Compared with terrestrial or seafloor applications, some additional challenges exist in the upper ocean domain, namely: (i) the non-static boundaries and surface features present in the environment (Blondeau-Patissier et al., 2014); (ii) the lower signal-to-noise ratio for signals arising from water masses; and (iii) the presence of surface effects such as sunlight and sky glint (Emberton et al., 2016). Studying ocean surface features for analysis of surface slicks, currents, fronts, waves, and wind interactions were discussed in Ryan et al. (2010) and Chen (2012), with an emphasis on synthetic aperture radar images. High-frequency (HF) radar was also combined with remote sensing in Das et al. (2010), where the sampling and hotspot prediction of harmful algal blooms is examined. This was also studied in Bernstein et al. (2013), where a shore-based recognition pipeline is suggested, building on remote sensing data for event detection, feature localization, and trajectory prediction. Frolov et al. (2013) analyzed the spatial and temporal decorrelation scales seen in marine algal blooms using fluorescence line height imagery, to strategize monitoring of such episodic events. Frolov et al. (2012) investigated short-term prediction of surface currents using HF radar observations to develop a linear autoregression model, where the climatology of conditional mean flow fields for upwelling, downwelling, and relaxation in Monterey Bay is presented. Optical water types were found using fuzzy clustering analysis on spectral information in Eleveld et al. (2017), aimed towards identifying different types of lakes, e.g., clear versus turbid waters. At large spatial scales, Oliver and Irwin (2008) used remote sensing data to automatically resolve different oceanic regions with certain spatiotemporal characteristics, to monitor the effect of El Niño events.

A number of approaches have been explored combining onboard models and remote sensing to guide sampling in the ocean. Smith et al. (2010) utilized forecasts from a high-resolution ocean model combined with remote sensing to pre-plan missions with multiple autonomous underwater vehicles (AUVs). Areas with high concentrations of chlorophyll a (a proxy for phytoplankton abundance) are identified from the satellite imagery and simulated forward in time using an ocean model. Presenting only simulated results, Smith et al. (2010) showed the potential and also the challenges of leveraging prior data. Issues related to small-scale discrepancy between model simulation (used for planning) and the actual conditions, aligns with the assumptions in our work, and the fact that we are focusing on predicting and planning based on large-scale (regional) features. Chao et al. (2017), provided a discussion and preliminary results in closing the loop between numerical ocean models, robotic platform sampling, and data assimilation. Multiple information streams are proposed to update and improve sampling strategies without human intervention. However, the robotic assets depend on human involvement and robust communications using shore-based assimilation and planning methods. Consequently, this drives the need towards elevated levels of autonomy and onboard sampling strategies for situational awareness, such as presented in our work.

This article describes a novel way of building compact ocean models using remote sensing products for use in robotic sampling. As our work uses pre-processed data (1 day average SST), some of the raw information (e.g., quality flags) are forsaken for practical purposes of obtaining and working with the data. The proposed method provides the ability to work directly with both temperature and spatial patterns, as this information is carried along throughout the analysis using a combined feature vector. In contrast to other work, the classification results are confirmed and verified by an independent marine data source (preceding wind history), allowing oceanographic processes to be tied to the subsequent analysis and use of the data. Wind drives horizontal and vertical circulations, which in turn determines magnitudes, gradients, and spatial patterns in SST (Rosenfeld et al., 1994). The method is unsupervised and can be used with a very small number of images (≤ 100).

3. Preliminaries

3.1. Satellite data sources

The SST remote sensing images used as a basis for this model are provided from a high-resolution radiometer onboard the Polar-orbiting Operational Environmental spacecraft (POEs) NOAA-17 and NOAA-18. The data are processed and mapped to an equal angle grid (0.0125
degrees latitude by 0.0125 degrees longitude) using a simple arithmetic mean, producing both individual and composite images from 1 to 14 days duration. This may provide some averaging artifacts, see Figure 15a. The nominal accuracy is about 0.7°C, covering the west coast of North America. 

We used 1-day average SST images, similar to Figure 3a, from 2015, 2016, and 2017 in this work covering Monterey Bay, California (marked in Figure 3b), for a 100-day period from March until the start of August, yielding 300 images in total. This period of the year was chosen in large part because wind-driven coastal upwelling and associated thermal signatures are strongest. Owing to cloud cover, only about 25% of the images had a quality that was useful for our application, giving us a total of 74 images to work with. Local cloud and fog cover can limit the data availability; henceforth, longer time periods should be considered for inclusion in the data sources if images are influenced by these factors. Equally, too much exposure can create data gaps that introduce bias into the final model, as images and their spatial patterns are left out of the analysis. Cloud and fog cover is further a motivating factor for actually deploying autonomous vehicles that can aid in estimating ocean conditions. Although the SST images in this analysis are limited to spring and summer, coastal upwelling circulation in this region continues into the fall season and is closely linked to coastal land features (Rosenfeld et al., 1994). Therefore, the interpretation of structural information by these methods should be similarly applicable during fall. However, the magnitude of SST shifts during fall due to seasonal warming throughout the region. This aspect of the environment motivates a seasonally dependent method, using images that are consistent seasonally.

The time stamp of images in Figure 2, shows availability over sequential days, and therefore the likelihood of having a similar mean temperature. However, SST patterns can be vastly different in sequential days because energetic currents change water mass distributions rapidly, motivating the case for looking at spatial similarities.

3.2. Unsupervised learning and classification of SST images

In order to quantify different spatial patterns in SST images, we use an unsupervised classification method based on sparse image representation. The compressed representations are obtained from employing dictionary learning techniques, before hierarchical clustering (Everitt et al., 2011) is performed in several steps to classify the images. The approach aims at classifying images with the same SST pattern into dominant/archetypical classes, having distinct oceanographic significance. Clustering the raw image data (pixels) is not effective, owing to their high dimensionality. Using bulk characteristics, such as the mean, min, and max temperature, is also possible. This approach can obtain good results, but leads to unnecessary smoothing/blur (in the conditional mean) as classes are combined without spatial information. To demonstrate this, a comparison of the classification variability is presented in Section 4.2. Consequently, clustering a compressed/sparse representation is more viable, as both temperature magnitude and spatial information can be combined together. Moreover, the dimensionality can be kept low, making it easier to cluster and hence identify and differentiate between the characteristics of each image.

3.3. Dictionary learning

Sparse dictionary learning (Aharon et al., 2006; Mairal et al., 2009) is a method that can be used to build sparse representations of data. The resulting format is a set of coefficients, collected as a sparse code. Dictionary learning is similar to principal component analysis (PCA) (Jolliffe, 2011) (also known as empirical orthogonal functions) in that the coefficients form a linear combination of certain basic elements, referred here as atoms. The atoms can be considered as instances of “characteristic patterns” that can be combined to reconstruct the input. The combined matrix of atoms is called a dictionary, usually denoted as \( D \), while the sparse codes are usually noted as \( \alpha \), having the relationship

\[
x_k = D\alpha_k
\]

where \( x_k \) is the raw image \( k \), with \( D \) as the dictionary, and \( \alpha_k \) as the unique coding for that specific image. We use MiniBatchDictionaryLearning or MBDL from Pedregosa et al. (2011) to find the dictionary using the least angle regression (LARS) method (Mairal et al., 2009), while the sparse codes are found using an orthogonal
matching pursuit (OMP) algorithm (Tropp and Gilbert, 2007) that greedily selects the dictionary atoms sequentially, through computation of the inner products between the image and dictionary columns. The loss function used in this optimization can be formulated as

\[
l(x, D) = \min_\alpha \frac{1}{2} \| x - D\alpha \|_2^2 + \lambda \| \alpha \|_1
\]

where \( \lambda \) is a regularization parameter (set to 1 in our implementation), \( \| \cdot \|_2^2 \) denotes the squared Euclidean norm, and \( \| \cdot \|_1 \) denotes the \( \ell_1 \) norm. This loss function, also known as basis pursuit (Chen et al., 2001), encodes two essential optimization criteria. The first term seeks to minimize the reconstruction error (depending on \( D \) and \( \alpha \)), while the second seeks to attain a sparse solution for \( \alpha \). As we also have to find \( D \), to solve for the function above, this is usually rewritten as a joint optimization problem with respect to the dictionary \( D \) and \( \alpha = [\alpha_1, \ldots, \alpha_n] \) as

\[
\min_{D, \alpha} \frac{1}{n} \sum_{k=1}^{n} \left( \frac{1}{2} \| x_k - D\alpha_k \|_2^2 + \lambda \| \alpha_k \|_1 \right)
\]

where \( n \) is the total number of images. To obtain the best dictionary \( D \) and sparse code \( \alpha \) the optimization alternates between optimizing one parameter while keeping one parameter fixed (see Mairal et al. (2009) for details). The images are decomposed into patches with a specific size, and fed into the algorithm in batches, to improve convergence performance. The number of patches is non-random and determined by the patch size, following a left to right, top to bottom extraction pattern. For the specific SST images used here, the patches also include a land/cloud mask. The size of these patches will influence the quality of the result and needs to be selected based on the small- and large-scale similarities in the input images; sensitivity analysis should therefore be used to identify this parameter (see Section 4.1). In the implementation used here, finding the dictionary \( D \) using the MBDL library involves an iterative process that uses a random state to initiate the model, hence reproducible results require using the same predefined random seed.

4. Methods

4.1. Proposed classification methodology

In separating and collecting dominant spatial patterns from SST images an important aspect that we emphasize in this work is that images sharing a common oceanographic evolution need to be identified. This implies finding images that share a mutual historic progression of, for example, wind and currents, that contribute to shaping a particular environmental condition. Involvement of local oceanographic expertise and knowledge is therefore essential in finding a separation scheme that is justifiable. This is especially true since the images are snapshots of a continuous process and separation into classes will imply some form of discretization.

The classification method builds off of the idea of classifying images based on a sparse representation (sparse code), instead of their high-dimensional pixel space. The classification is also conducted in two steps, illustrated in Figure 4 as a branching graph with a new dictionary and
sparse coding generated after each step (i.e., “initial classification” and “secondary classification”).

This two-step classification increases the performance of the method, as the larger variability (magnitude) is handled in the first classification, followed by a more fine scale separation (spatial patterns) in the second. The detailed steps involved can be described as follows, for image $k$ and each class $j = 1, \ldots, m$.

1. Normalize all images with the global mean and standard deviation.
2. Extract a fixed number of patches from the SST image (e.g., 260 patches of size $40 \times 40$ pixels), and stack these in a one-dimensional (1D) vector.
3. The 1D vectors are collected in the data matrix and fed into the MBDL subroutine to retrieve the initial dictionary $D_j$.
4. Each SST image is then compressed using the dictionary $D_j$ to form a unique sparse coding for the image $\alpha^k_j$.
5. These first $\alpha^k_j$ coefficients are then classified using hierarchical clustering.
6. Steps 2 and 3 are then repeated for the images within the derived classes to retrieve a new sub-class specific dictionary $\tilde{D}_j$.
7. Step 4 is repeated using the sub-class-specific dictionary $\tilde{D}_j$ to obtain the final sparse codes $\tilde{\alpha}^k_j$.
8. The final sparse coding $\tilde{\alpha}^k_j$ is now further classified into sub-groups with a higher level of similarity using another step of hierarchical clustering. This is the secondary classification as shown in Figure 4. This final classification is subject to a criteria where the mean number of images across the final classes should be at least 3.0 (see below); if not, the images are flagged as being too distinct.

It is important to ensure that the final classification result is not too distinct, e.g., having only one image in a class, resulting in loss of effectiveness towards building a compact model. It is, therefore, important to consider the separation distance (SD) in hierarchical clustering as a measure of similarity of classified images. A practical aspect of using hierarchical clustering is that compared with other algorithms such as $k$-means (MacQueen, 1967), where the number of clusters must be pre-specified, hierarchical clustering can use the dendrogram to generate the clustering structure. Consider therefore the dendrogram in Figure 5 used for finding SD in the initial step. Lowering the SD yields more classes with fewer images and vice versa. The distance metric used for clustering is the Ward sum-of-squares minimization metric (Ward, 1963), chosen since it is more permissive of cluster shape/size assumptions (Anderberg, 1973), and which performed better on the sparse codes. In using this distance measure, it is important that the data vectors we are operating on are normalized by the mean and standard deviation (Wilks, 2011), hence Step 1 above. The SD therefore, for the initial clustering is found based on the dendrogram and evaluation of the separation results. As noted in Wilks (2011), setting the SD usually requires a subjective choice that depends on the goal of the classification. There exist various statistical machine learning tools that consider the bias versus variance trade-offs (e.g. James et al., 2013), but there is no single “best” approach. As this initial grouping is followed by a secondary classification, this choice is not as decisive a factor, as the secondary clustering, as only the major features are to be identified. Nevertheless, local oceanographic expertise and knowledge should be involved to find a justifiable separation scheme. For the first step of the classification, a general rule of thumb would be to avoid a SD that yields too sparse a set of classes.

Fig. 4. Based on the dictionary $D$ a sparse code $\alpha$ is found. These codes are used to classify images into distinct classes. After an initial classification this step is repeated, with a new dictionary and codes generated from the images within a class; branching into smaller sub-classes with higher levels of similarity.

Fig. 5. The dendrogram for the initial clustering step. Four classes are being separated using a distance measure set to 700. The numbers on the x-axis are either image index or cluster size (in parentheses).
The patch size is adjusted to yield different results. The relationship between the patch pixel size and may lead to degraded classification accuracy for patch sizes above $30 \times 30$, without any gain in accuracy. Based on this analysis we use a patch size of $40 \times 40$ pixels. Figure 6b shows that a dictionary size of 8 as the number of images being classified are fewer and their similarity is higher (filtered by the initial classification), hence a more descriptive code can be applied.

### 4.2. Comparative analysis

As the proposed method provides enhancement of classification by including structural information, the performance of the methodology is compared by using (i) hierarchical clustering of temperature metrics only (minimum, maximum, min–max range, and mean temperature, i.e., no sparse codes), and (ii) directly classifying the raw images using a $k$-means approach. To deal with an unknown number of clusters, we use $X$-means (Pelleg and Moore, 2000), while for the hierarchical approach we use the dendrogram. As the temperature magnitudes from the SST can contain several strong separating factors some initial categorization can be achieved; however, neglecting spatial information leads to unnecessary variability. By comparing the ICV given in (4) between the proposed, temperature-only, and $X$-means approach derived classes, the effect of using spatial structure can be demonstrated. This is shown in Figure 7 where a histogram presents the ICV for the derived classes, where each unit block is a class with a corresponding ICV.

$$\mu_{ij} = \frac{1}{n_j} \sum_{k \in \Theta_j} x_{ik}$$

where $x_{ik}$ is the temperature at location $i$ in image $k$, $\Theta_j$ is the set of images belonging to class $j$, and $n_j$ is the number of images in this set. Moreover, $\mu_{ij}$ and $\sigma_{ij}^2$ are the sample mean and variance in temperature at location $i$ for class $j$. Using ICV as a metric, the patch and dictionary size is chosen based on the settings that give the lowest ICV after several iterations for each setting with a non-constant random seed. The patch size sensitivity (Figure 6a), shows a drop in ICV with increasing patch size, that gradually flattens as the patch size increases. The variability is slightly increasing for patch sizes above $30 \times 30$, without any gain in accuracy. Based on this analysis we use a patch size of $40 \times 40$ pixels. Figure 6b shows that a dictionary size of 4 is sufficient to achieve both low ICV spread and value. For secondary classification, the dictionary size is increased to 8 as the number of images being classified are fewer and their similarity is higher (filtered by the initial classification), hence a more descriptive code can be applied.

For the secondary clustering step, finding a SD that does not result in an imbalance in variance (i.e., too big/small classes), becomes important. The goal here is to find similar images that can be used in a conditional mean, while also excluding images that are too distinct. Similarity is, as noted, controlled by the SD. To achieve an adequate final result we iteratively change the SD until we reach a classification result where the mean number of images across the sub-classes is above or equal to a certain threshold, set to 3.0. This threshold is chosen such that the generated classes must span several images and avoid creating classes with dissimilar magnitude and spatial structure. For example, if a secondary classification with 12 images starts with a SD = 500 and yields the sub-classes $[1, 2, 3, 4, 5]$ with associated image counts of $[5, 4, 1, 1, 1]$, the mean number of images in each class equals 2.4, and three classes have only one image. This does not satisfy our criterion, and the SD can therefore be increased, relaxing the measure of how similar the images need to be. The SD is then relaxed (increased) until one of the classes containing only one image is combined into one of the other five, as this reaches our criterion.

Parameters that are associated with the above procedure, are shown in Table 1. The patch and dictionary size can be adjusted to yield different results. The relationship between the patch size $x_{pa}$, dictionary size $x_{d}$, and code length $x_{cl}$ is given, using the fixed dimensions $52 \times 59$ for the images, as $x_{cl} = ((52 - x_{pa} + 1) \times (59 - x_{pa} + 1)) \times x_{d}$ and, therefore, the code length depends on the square of the patch size, while increasing linearly with dictionary size. This code length is important as clustering becomes more complex and may lead to degraded classification accuracy for higher-dimensional vectors, as can be seen in Figure 6a. Parameter selection was done by comparing the intra-class variability (ICV) for each class $j = 1, \ldots, m$ measuring how different the images within a class are by comparing the temperature variability at each location across the images, given as

$$ICV_j = \sum_i \sigma_{ij}^2,$$

$$\sigma_{ij}^2 = \frac{1}{n_j} \sum_{k \in \Theta_j} (x_{ik} - \mu_{ij})^2,$$

which can be achieved; however, neglecting spatial information leads to unnecessary variability. By comparing the ICV given in (4) between the proposed, temperature-only, and $X$-means approach derived classes, the effect of using spatial structure can be demonstrated. This is shown in Figure 7 where a histogram presents the ICV for the derived classes, where each unit block is a class with a corresponding ICV.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input dataset</td>
<td>74 images, size: $52 \times 59$ pixels</td>
<td>Type: 1 km, 1-day average SST</td>
</tr>
<tr>
<td>Patch size $x_{pa}$</td>
<td>$40 \times 40$</td>
<td>The patch pixel size</td>
</tr>
<tr>
<td>Sample count</td>
<td>260</td>
<td>The number of sample patches from each image</td>
</tr>
<tr>
<td>Dictionary size $x_d$</td>
<td>4</td>
<td>Number of atoms in the initial dictionary</td>
</tr>
<tr>
<td>Dictionary size $x_d$</td>
<td>8</td>
<td>Number of atoms in the secondary dictionary</td>
</tr>
<tr>
<td>Distance metric</td>
<td>Ward</td>
<td>The distance metric used by the hierarchical clustering</td>
</tr>
</tbody>
</table>

### Table 1. Overview of raw data, as well as parameters used in dictionary learning and hierarchical clustering.
The figure shows that classifying using only temperature metrics or raw data leads to higher variance. As the sparse codes are able to compress information about temperature and spatial patterns into one vector (with lower dimension), the resulting evaluation in the downstream classification task is simplified; this, in turn, leads to a more accurate classification with lower variance. Not surprisingly, X-means clustering has the worst performance, as the method only splits the data into two classes (cold and warm). Classifying the data further using X-means results in no further separation.

5. Classification results

The final classification groups are used to make different conditional means that together constitute a compact model. An important aspect of the approach is to verify that images that are classified together share a common formation history, i.e., having been influenced by the same sequence of physical processes. Having a shared evolution of oceanographic conditions also implies that there exists a common and distinct spatial pattern that can be clustered together to make what we choose to call a condition. Wind observations are effective for this purpose as it is the dominant driver of circulation, SST variability, and environmental structure in this region of study. The use of other sources of data, rather than wind, is also possible and will vary depending on the study region. The comparison and verification is shown in Figure 8 together with the initial clustering result.

5.1. Initial classification

The initial classification is dominated by the temperature magnitude, as this is a stronger separating factor than spatial patterns. The SD used in hierarchical clustering was found using the dendrogram in Figure 5, producing four classes. Figure 8 shows these together with the aggregated 4-day preceding wind history. The images appear in arrangement going from cold to warm, or in the oceanographic context, from upwelled to relaxation dominated waters. The associated winds confirm this by showing a correspondence between the upwelled (class 1, cold) SST images with strong north-westerly winds, and weaker more spread wind pattern for the relaxation (class 4, warm) SST images. The intermediate classes comprise images covering the transition between these two oceanographic conditions. The initial classification has two classes of mainly cold water ([class 1 and 2] \(10 + 27 = 37\) images), and two classes of warmer waters ([class 3 and 4] \(15 + 22 = 37\) images).
images), with the same number of images. This suggests a balance between upwelling and relaxation events.

Figure 9 shows the average $u$ and $v$ components of the wind history for each class. Similar to Figure 8, the average trend shows class 1 has the strongest average winds, and class 4 the weakest. A strongly negative $v$ component of the wind is equivalent to north-westerly, i.e., blowing from the northwest and upwelling favorable. There is also a strong diurnal signal in the winds, evident in all classes in the $u$ component. This predominantly east–west “sea-breeze” is driven by differential heating of land and ocean through day/night.

The initial classification shows that some structure follows from the temperature, but as seen in Section 4.2, using temperature alone will not achieve the best clustering result. Magnitude information helps us avoid a pitfall where images that have a similar pattern but different temperature magnitude are clustered together, disrupting the conditional means by averaging images that represent different oceanographic states.

To understand this initial classification further Figure 11 shows the projection of the sparse codes using PCA projected onto a 2D plane, visualizing the information contained in the vectors. As noted previously, as the SST images are covering a continuous process some images are overlapping and could potentially be associated with more than one class. It is also possible to identify certain images that can be deemed as outliers. The colors in Figure 11a show where the hierarchical clustering distance, $SD$, is making the distinction.

5.2. Secondary classification refinement

The images in each class are to be further distinguished in the secondary classification. The images in this step are now already sorted according to matching temperature, which we have shown, by comparing the wind history, can be traced to distinct evolution of oceanographic conditions. Within each class, a new dictionary $\tilde{D}_j$ can now be found,
that can specialize in finding a final sparse code $\tilde{a}_k$ that factors in more spatial information, as shown in Figure 7. The parameters for the secondary classification, found in Table 1, are similar to the initial classification apart from a larger dictionary size. As noted, the dictionary size has increased because the number of images being classified are fewer and their similarity is higher in this step, hence a more descriptive code can be applied.

Figure 10 shows the resulting secondary classification, with a total of 19 sub-classes (see Table 2). The classes containing only one image are deemed too distinct for further inclusion into the compact model. As expected the mean class size is above 3 as specified from the criteria in Step 8, with ICV values from 0.09 – 0.13. It is also apparent that images are now sorted both by temperature as well as spatial features (see e.g. the upper left corner of Figure 10c). However, cloud cover can pose challenges. In some cases, this “false” pattern match can be such a dominating feature that images with a similar false feature are classified together, (e.g. the lower group in Figure 10a) as shown in Figure 12.

Taking the average across each sub-class in Figure 10 yields the different conditional means and the final compact model, shown in Figure 13b. Rather than relying on the means from the initial classification in Figure 13a, these representations hold more spatial information that is advantageous when trying to predict the current state of the environment. From these conditional means one can observe not only the transition from cold to warm conditions, but also the varying spatial structure that develops within and outside Monterey Bay.

6. Forecast and sampling policies

We now use these conditional means as a compact model from which a prediction of the environment, specifically the SST field, can be made. Having the capability to compare in situ data with the model provides a way to determine which types of historical conditions that best fit these data. On this basis an estimate of the current state can be formed, usually taken as a weighted combination of the “best” class candidates. A common method for verifying the reliability of the prediction is to compare the root-mean-square error (RMSE) of the class mean with the average class spread, as suggested in Fortin et al. (2014); the idea is that the standard deviation of the class spread should be approximately equal to the RMSE. We further study how the prediction probabilities of different states depend on the data, and how this can be used in designing sampling strategies.

6.1. Predicting the environment

From Figure 1, an intuitive way to predicting the environment is by combining one or several classes using a weighted average. During robotic deployments, as with a WaveGlider ASV, the prediction is made conditional on data $y = (y_1, \ldots, y_{np})$, where $np$ is the number of measurements made during the survey. Based on satellite data, the

Table 2. Classification results. The ICV, given in (4), measures the mean sub-class variability. Asterisks (*) mark groups with only one image, which are considered too distinct and hence removed.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Classes</th>
<th>Image count</th>
<th>Mean class size</th>
<th>ICV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>[1 2 3 4]</td>
<td>[22 10 27 15]</td>
<td>18.5</td>
<td>0.30</td>
</tr>
<tr>
<td>2nd, Class 1</td>
<td>[1 2 3]</td>
<td>[3 2 5]</td>
<td>3.33</td>
<td>0.12</td>
</tr>
<tr>
<td>2nd, Class 2</td>
<td>[1 2 3 4 5 6 7*]</td>
<td>[2 3 5 2 7 1*]</td>
<td>3.14</td>
<td>0.09</td>
</tr>
<tr>
<td>2nd, Class 3</td>
<td>[1 2 3 4]</td>
<td>[2 2 2 9]</td>
<td>3.75</td>
<td>0.13</td>
</tr>
<tr>
<td>2nd, Class 4</td>
<td>[1 2* 3* 4 5 6 7 8]</td>
<td>[2 1* 1* 4 4 2 8 5]</td>
<td>3.37</td>
<td>0.13</td>
</tr>
</tbody>
</table>
The probability \( P(\text{class} = j) \) is estimated as the fraction of images in each class over the total number of images, \( P(\text{class} = j) = \frac{n_j}{n} \). Conditional on data \( y \) the probability of class \( j \) becomes

\[
P(\text{class} = j|y) = \frac{p(y|j)P(\text{class} = j)}{p(y)}
\]

The class likelihood, in (5), is approximated as a multivariate normal distribution:

\[
p(y|j) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}} e^{-\frac{1}{2}(y - \mu_j)^T \Sigma_j^{-1}(y - \mu_j)}
\]

where \( \mu_j = (\mu_{1j}, \ldots, \mu_{nj}) \) is the class \( j \) vector of mean values along the survey trajectory, and \( \Sigma_j \) is the associated covariance matrix at these \( n_p \) sampling locations. This covariance is defined by elements \( \Sigma_j(i, i') = \text{diag}(\sigma_{ij}) R \text{diag}(\sigma_{ij}) \), where \( R \) is a distance-based correlation matrix (Matérn, 2013), i.e., \( R(i, i') = (1 + \phi h_{ij}) e^{-\phi h_{ij}} \), where \( h_{ij} \) is the Euclidean distance between sampling locations \( i \) and \( i' \), and \( \phi \) is indicative of the correlation range. In this paper we use \( \sim 15 \) km, based on Frolov et al. (2014: Figure 4b) showing decorrelation scales for Monterey Bay.

Prediction is made using class probabilities as a function of the data gathering window. This means that we integrate one more observation at every step, and re-calculate the probability over the classes, given this growing subset of data. Owing to the spatial correlation in the model, induced via \( R \), the assimilation of one more observation will not have the same effect as it would for independent data. The final prediction at location \( i \), based on available data \( y_i \), is a weighted average according to the conditional distribution in Equation (5),
\[
\tilde{SST}_i = \sum_{j=1}^{m} P(\text{class} = j | y) \mu_{ij} \tag{7}
\]

where \( \mu_{ij} \) is the conditional mean in class \( j \). A brief example using this approach is presented using data from a WaveGlider ASV. This vehicle records the temperature at 0.4 m depth using a Seabird CTD (conductivity, temperature, and depth) sensor, which we will use to predict the SST.

We use data from the 17th May 2018 in Monterey Bay, during the CANON field experiment. The prediction, using (7), is compared with the actual 1-day average SST for that day. Figure 14 shows the recorded temperature profile across the survey locations (the survey path is shown in Figure 15a) together with the corresponding profile from classes 0, 8, 10, and 13. Clearly, some classes match the data better than others. Classes 8 and 10 follow closely throughout, whereas classes 0 and 13 have poor correspondence. We use this data to predict the environment using \( P(\text{class} = j | y) \) as the weights. Thus, the likelihood is therefore expected to be high around classes 8 and 10, and low for the others. The prediction is shown in Figure 15b, together with the actual SST image in Figure 15a. The accuracy of the prediction is, as expected, better on a larger scale, with some smaller spatial features that are not captured by the model. Figure 15d shows this temperature difference spatially. Note that the nominal accuracy of the SST images is about 0.7°C. As the original daily composite SST image is not without error, e.g., noise from daily averaging, some regions will show exaggerated difference. It is also reasonable to assume that there is a discrepancy between the WaveGlider data and the daily average SST, which also contributes towards estimation error. This can be seen with closer inspection of Figure 15a, as the track and overlaid temperature are colder than the average SST.

\[
\text{RMSE}(j, y) = \sqrt{\frac{\sum_{i=1}^{n_p} (y_i - \mu_{ij})^2}{n_p}} \tag{8}
\]

There is a correspondence between the likelihood \( P(\text{class} = j | y) \) and the RMSE. As expected the maximum likelihood and the lowest RMSE occur around class 10. As the likelihood adjusts for spatial correlation and prior probability \( P(\text{class} = j) \), there is some difference between the two measures of similarity (e.g., class 9). Comparing the mean RMSE of the estimated SST against the class spread (0.45 versus 0.77), indicate that we are in accordance with the reliability measure discussed in the beginning of this section.

### 6.2. Evaluation of model prediction error

For discrete models, such as that developed here or by Lilleborge et al. (2016), the intent of data collection is to pull the predictive probabilities closer to 0 or 1 (Eidsvik et al., 2015). The \textit{a priori} prediction error (before any data are recorded) is given via the most likely ocean state class as \( j^* = \arg\max_j \{P(\text{class} = j)\} \) as

\[
PE = 1 - P(\text{class} = j^*) \tag{9}
\]

Conditional on data \( y \) the prediction error is \( PE(y) = 1 - P(\text{class} = j^*(y) | y) \), where now the most likely
ocean state is given as \( j^*(y) = \arg\max_j \{ P(\text{class} = j | y) \} \), with probabilities defined in (5).

To evaluate the model and different sampling strategies before any data \( y \) is collected, we can look at the average posterior prediction error obtained by integrating over all possible data as

\[
PE(y) = E \{ 1 - P(\text{class} = j | y) \} = \int (1 - P(\text{class} = j^*(y) | y)p(y)) dy \tag{10}
\]

where \( j^*(y) \) is the class with the largest probabilities conditional on \( y \). The improvement made by data collection can now be compared over various experimental designs (i.e., survey paths) using the Monte Carlo approach outlined in Eidsvik et al. (2015: Ch. 5.6). To analyze the effect of data gathering, we start by generating synthetic data from the model. This entails drawing a random class \( j^b \) from probabilities \( P(\text{class} = j), j = 1, \ldots, m \), and drawing data \( y^b \) conditional on this class based on the likelihood model in (6). We set

\[
y^b = \mu_j + L_j z
\tag{11}
\]

This approach uses the Cholesky factorization (Nash, 1990) of the covariance matrix \( \Sigma_j = L_j L_j^T \) along with a length \( n_p \) vector \( z \) of independent \( N(0, 1) \) variables. Examples of generated data can be seen in Figure 16. The final error prediction can now be computed using the generated data in a Monte Carlo approximation:

\[
PE(y) \sim \frac{1}{B} \sum_{b=1}^{B} (1 - p(j^*(y^b) | y^b)) \tag{12}
\]

where \( B \) is the number of iterations (we used \( B = 500 \)).

It is now possible for a practical comparison of the reduction of prediction error for different survey times (the effect of gathering more data) and survey locations (survey trajectories). The results are shown in Figure 17, as a function over the survey duration, and for three different survey paths, using the WaveGlider and assuming a platform speed of approximately 2.2 knots. Using this, different survey lengths can be correlated to mission time.

Figure 17a shows the prediction error for a given mission duration. Naturally, longer missions result in more observations and less error, reflecting the amount of data that is available. The effect of different survey paths is shown by using three different routes, shown together in Figure 17b, with their evolution and final \( PE(y) \), as shown in Figure 17a. Path 1, which crosses both the inner and outer bay, produces the lowest prediction error \( (PE(y) \sim 0.15) \).

Fig. 16. An example of 50 synthetic survey lines data for class \( j = 12 \) and \( j = 3 \), with missions lasting 16 hours.

Fig. 17. (a) Predicted error versus mission time. The effect of gathering more data (expressed as mission duration) on the prediction error, calculated for each of the three paths. The error drops (from close to the prior probability \( PE \sim 0.64 \) as more information is obtained. (b) The different mission paths. By comparing different survey paths, one can observe that some locations are more informative than others, showing difference in \( PE \) curves.
A possible reason for this is that path 1 crosses both gradients inside the bay, as well as gradients that are prominent further offshore, covering the usual band where upwelling fronts occur. The curve starts at the prior probability $PE^{-0.64}$, which arises from (9), estimated as the fraction of images in each scenario over the total number of images. This type of investigation is useful, since it provides an estimate of both the value of mission duration and location, which is often an unknown when planning survey campaigns. Such an analysis can also be conducted across multiple platforms; each platform can be evaluated by simulating a different coverage (survey speed), spatial correlation ($\theta$), measurement noise, etc. Optimization of coverage versus cost is also possible, finding effective solutions that maximize the cost per observation. Such methods of reasoning (“What is the value of the data and how much data is enough?”) are often referred to as value of information analysis (Eidsvik et al., 2015).

7. Discussion

In the absence of remote sensing, the description of regional high-resolution data may be unavailable. A compact model, such as that developed in this work, and the supporting statistical tools can help provide contextual low-resolution information, by using in situ observations. Reducing the global uncertainty is necessary for enabling efficient planning of vehicle surveys, that rely on evaluating the conditions at unexplored locations, as well as variability and associated correlation structures. With this in mind, some aspects of the presented approach are discussed, in order to shed light on potential benefits and pitfalls.

Assimilation using an onboard numerical ocean model that accounts for time is currently not possible or practical owing to time and computational limitations, hence compact or reduced order models are needed. The current compact model is static, i.e., the classes themselves are not modified during the mission. In practice, this means that small scale features will not be well resolved. One could use Gaussian process regression (Rasmussen and Williams, 2006) to assimilate the observations and correct this locally. However, the primary capability here is to predict regional features, hence updating each class locally is of limited value, as it is the prediction at unobserved locations that are most interesting towards future sampling. Prediction works by taking a weighted average, using a likelihood function. There exist several strategies for finding an alternative weighting scheme. The current approach can be improved by including covariates (e.g., wind measurements) to further determine some of the global conditions and find the weights conditioned on this. The weightings can also be found using optimization such as sequential least squares programming (SLSQP) (Nocedal and Wright, 2000) to minimize the error between the observations and a weighted combination of the classes.

To create sparse feature vectors of the images that is suited for subsequent classification, we make use of dictionary learning techniques, as discussed in Section 3.3. Alternative methods using state-of-the-art deep learning approaches such as convolutional auto-encoders (Aljalbout et al., 2018; Song et al., 2013) can also be used to create similar sparse representations without labeled data. However, given the limited size of the dataset (74 images) these approaches are not feasible to use, even with data augmentation, as they require larger datasets (e.g., LeCun et al., 2010) in order to effectively tune the large number of parameters in the network; this is well-known issue for neural network architectures (Liu et al., 2017). In remote sensing applications, increasing the size of the dataset is also not straightforward (Mountrakis et al., 2011), as discussed in the following in more detail.

There are inherent limitations in using SST satellite images owing to cloud cover, such as the lack of uniformity in data acquisition (gaps in data), and the introduction of artifacts (e.g., interpolation around missing pixels). Thus, images have to be quality controlled before inclusion into the dataset, which may end up becoming smaller than expected. One question, in this regard, is whether numerical ocean models could be used instead. Ocean models can offer repeated and synoptic fields at different scales and resolutions, including surface and subsurface patterns. However, at the current skill level, coastal ocean variability and structure cannot be estimated on scales and accuracies sufficient for definite representation (Lermusiaux, 2006), i.e., we cannot expect the models to tell us right where a filament, eddy, or bloom will be located, as an example. Assimilation of SST and HF radar current data (see, e.g., Frolov et al., 2012) can certainly help the models move closer to ground-truth, but the nudging of 3D fields with spotty 2D data can create its own issues. For instance, spotty coverage in observations strongly limits how effectively a model can be corrected toward observed patterns, and this data assimilation can create artifacts that can be misinterpreted as “features of interest.” Ultimately, effective observation of the ocean will require a joint effort between a range of data sources operating at different scales, where assimilation of in situ data collected from autonomous robotic platforms will be crucial. Methods, such as that suggested in this article, will be an important tool to strategize this sampling.

An interesting application of this work, once a model is built, is to calculate a vehicle trajectory that results in the greatest level of class separation, that is, a path that best reduces the uncertainty in the model by fast convergence to a ranking (a probable scenario); similar to the analysis in Figure 17a. Finding such a route can improve prediction of the oceanographic state by gathering temperature data in areas that will enhance discrimination, and can locate areas where multidisciplinary sensing from platforms will be most informative. Criteria such as expected reduction of
variance, mutual information, or entropy can be used to find these locations.

Regional factors such as wind, bathymetry, and currents contribute to shaping the spatial patterns used for classification, which contribute to shaping SD and other parameters. Local oceanographic expertise is therefore necessary for validation. The method also involves some work related to tuning of hyperparameters, that will need supervision by oceanographers to ensure that a physically sound separation is used for building the model. Once configured, the method can operate on its own, and be automatically set to ingest new SST images. We use different settings for the dictionary size in the initial and secondary classification. This increases the length of the sparse codes, but allows for more spatial/pattern-based information to be used in classification; initially this might not be necessary as the temperature dominates.

8. Conclusion

We have developed a new methodology for classifying remote sensing products such as SST towards building compact models that can be utilized by autonomous vehicles to provide environmental estimates. The method allows continuous processes to be segmented and compressed onto a basis of classes/scenarios that can be used as a framework for informing trajectory planning and sampling design. This approach can enhance the effectiveness of ocean observing campaigns and in the end help scientists understand the influence of regional oceanographic variables. We show examples using real data from Monterey Bay, California, where the compact model is combined with in situ data to predict regional oceanographic states and bulk features. The results show that local observations can be used to yield information on a synoptic scale, but are limited to only resolving regional details. We have also evaluated the prediction error of the model and demonstrated the sensitivity to data, from the perspective of an ASV.

Acknowledgements

The WaveGlider data was provided from the MBARI CANON³ campaign with support from Christopher Wahl, Devon Northcott, and Francisco Chavez. The authors are also grateful for the support from MBARI engineering, science, and marine operations. The work was conducted when TOF was a visiting scholar at MBARI.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was part of the ENTiCE project (http://sintef.no/en/projects/entice/) funded by the Research Council of Norway (project number 255303/E40), the Nansen Legacy Program (project number 276730), and AMOS (https://www.ntnu.edu/amos), Center of Excellence, project number 223254. TM acknowledges support from the Stanford Center for Earth Resources Forecasting and from the Dean of the Stanford School of Earth, Energy, and Environmental Sciences. MBARI authors are funded by the David and Lucile Packard Foundation.

Notes

1. The data are publicly available through the National Oceanic and Atmospheric Administration (NOAA) NWS Monterey Regional Forecast Office and the CoastWatch program, from their ERDDAP server https://bit.ly/2ngyP6c.
2. The u signifies the zonal velocity component of the wind, while v the meridional component.

References


