

Interactive comment on “Using Canonical Correlation Analysis to produce dynamically-based highly-efficient statistical observation operators” by Eric Jansen et al.

Anonymous Referee #1

Received and published: 19 February 2019

General comments

The paper describes a method for producing efficient estimates of skin and subskin SST using the Canonical Correlation Analysis (CCA) statistical technique. This is described in the context of observation operators which are used in data assimilation schemes to provide information about the model-observation misfits in order to correct the model. The aim of the paper is to make better use of satellite SST observations in data assimilation systems compared to most existing operational systems by improving the way the diurnal cycle of SST is represented. The work is therefore valuable and the paper is well written.

Near the end of the last section, there is a reference to another paper submitted by Korres et al. which appears to address a very similar topic by the same list of authors. This other paper should be referred to in the introduction and the aims of the two papers differentiated clearly. It might not be possible to address some of the comments below without using results presented in the Korres et al. paper. If that is the case perhaps the authors should consider merging the papers or having a two part paper.

The SOSSTA project is presented in three separate papers. The first (Pimentel et al. 2019) describes the modelling of the diurnal cycle in the Mediterranean Sea. The second (this paper) describes the method of building an observation operator by parameterising the results of an external model. It uses the dataset from the first paper as an example to demonstrate the method. The third paper (Korres et al., 2019) uses the GOTM datasets and the method presented in this paper and applies it in the POSEIDON data assimilation system. We believe that the topics are sufficiently different and self-contained to warrant publish them as separate papers. Moreover, we are hoping to publish additional papers such as Korres et al., 2019 that document the application of the SOSSTA operator in other models/systems.

Specific comments

1. The paper describes the method in a general sense with SST as a “use case”. It might help the flow of the paper if the use of the method for SST was more central to the paper so that, for instance, the introduction would have more information about the literature on SST analysis and data assimilation. The way it is currently structured, more review is needed on the various other types of meteorological and oceanographic observation operators used, for instance radiative transfer models.

The technique for building a parameterisation using canonical correlation analysis on an existing dataset is very general and can be applied in many situations. To emphasise the general nature of the method we decided to first describe the method and then focus on one specific application.

2. The scheme is not tested in a data assimilation system (although that appears to be done in the Korres et al. paper). Would the aim of a scheme using the CCA method be to correct for errors in the diurnal cycle of SST at 1 m depth using all available satellite SST measurements, or to correct the model's foundation SST?

The aim of the observation operator is to provide the best possible estimate of SST from the model background, based on the actual atmospheric conditions. The correction to the model temperature depends somewhat on the modelling of the model covariance in the data assimilation system. In general the correction will affect all the layers that in turn are affected by diurnal variability

3. There are shortcomings in the design of the experiments to test the performance of the CCA method in section 4.3. The validation is performed over the same period as the model data used to generate the CCA OO, and then compared to the same model data. There is no comparison of the CCA results to real observations.

The purpose of Sect. 4.3 is to show how the simplified parameterisation of the CCA method is able to approximate the full GOTM model. As the training dataset contains $O(1000)$ profiles and the parameterisation has only $O(10)$ parameters, there is no risk of overfitting.

Nevertheless, we agree that a demonstration of the method on a separate dataset would make the message of Sect. 4.3 stronger and remove doubts that people may have about the method. We have decided to revise Sect. 4.3 and to use an independent dataset for the comparisons. This independent dataset is obtained by withholding every other profile (along the zonal direction) in the input dataset from the calculation of the CCA OO.

4. The comparisons with other SST assimilation methods described in section 5 assume that the GOTM model and the atmospheric forcing driving it are correct. If the skin model in GOTM had a bias for instance, could it be worse to use the CCA OO based on that model than the other methods tested? A more independent way of assessing the CCA OO method based on GOTM is needed, e.g. by comparing to real observations.

This is a valid point and we fully agree. We have revised Sect. 5 of the paper and the comparison with other methods is now performed using the SEVIRI L3C dataset of subskin temperature. The performance metrics are calculated using only the withheld profiles, as described above at point 3.

5. There is not any discussion of the need for the adjoint of the observation operator in data assimilation (which is obviously very efficient when using the

CCA method).

We have rephrased the final comment of Sect. 3 to be more inclusive of 4D variational schemes: "The CCA OO is straightforward to implement in this scheme, since for H' and its adjoint $H'T$ it follows that: $H' = MT$, $H'T = M$ "

6. The GOTM dataset used for training the CCA is a model and not based on observations. The work of Pimentel et al. 2018 describes how the model represents the skin and subskin SST, but some more information is required here to justify its use. A brief description of how the model has been made to represent the skin and subskin would help (the highest level of 1.5 cm is not at the same depth as the skin or subskin). A summary of the assessment of the model compared with real observations is also needed here, otherwise there is no link to the real world.

We have expanded Sect. 4.1 and included the following: "The subskin SST represents the temperature at the base of the conductive laminar sub-layer of the ocean surface; for practical purposes it is represented by the temperature of the top model layer of GOTM (1.5 cm). The conductive sub-layer of the air-sea interface, associated with the cool-skin effect, is parameterised and dynamically computed within GOTM to produce a modelled skin SST. Further details are provided in Pimentel et al., (2019)"

Technical comments

1. Pg 2, line 5. The wording "are not or not sufficiently modelled" is a bit confusing on first reading. This lack of process representation in the model is often included in the representation error in data assimilation systems. A discussion on the relationship between the complexity of the observation operator, and the inclusion of representation error in the R matrix would be good here.

We did not include a discussion of the representation error as we felt that this would be beyond the scope of this paper. You are right that missing or insufficiently modelled processes should be taken into account in the covariance matrix. However, increasing the uncertainty is not an alternative to actually modelling the missing processes. This is especially true for processes that are not random but that create strong systematic biases, such as the diurnal cycle of SST that we discuss in Sect. 4.

We have included the following paragraph in Sect. 4: "Errors due to e.g. limited spatial resolution or unrepresented processes are generally included in the representation error. Representation errors have been extensively discussed within ocean applications (Oke and Sakov, 2008; Janjic et al., 2018). However, the diurnal variability of skin SST represents a potentially systematic error that requires a proper treatment rather than just the increase the representation component of the observational error."

2. Pg 2, line 6 -7. Does the cost of the “second” model depend on the observation number as implied here, or the (horizontal) model grid size? The cost of these models, e.g. a diurnal model, is cited as one of the justifications for implementing the CCA method. It is not obvious that a simple diurnal model is that expensive compared to the cost of the full GCM.

The use of an observation operator scales with the number of observations to be assimilated. When talking about the computational cost we are referring to state of the art diurnal models such as the presently used GOTM. Of course simpler models/parameterisations exist that are less costly to run, in fact the parameterisation used in the SOSSTA operator could be regarded as such a model.

3. Pg2, line 7. “needs” to “need”.

Thank you, it has been corrected.

4. Pg 3, eq (3). Normally matrices would be in uppercase but here you start using lowercase letters. This is particularly confusing when you use uppercase and lowercase of the same letters (e.g. u and v).

Unfortunately, the common notation for canonical coordinates and for singular value decomposition both use the letter v. We understand that this may be confusing and have decided to rename the canonical variables F and G. This removes the double usage of letters and allows us to capitalise all matrices.

5. What are the implications of eq (11)? It is taking into account the biases in the training observations and model. These presumably are not constant in time so how can this be applied in practice? There is no description of how these values are calculated in section 4, or their magnitude.

As the aim of CCA is to find correlations between datasets, it only considers variations of the variables with respect to their mean value. The CCA procedure subtracts the mean of each level, so the matrix M by itself would only relate temperature anomalies to each other. The offset factor K adds the mean values in order to relate temperature values instead of anomalies. It does not represent a form of bias, even in a perfect world K is not expected to be 0. The calculation is done according to Eq. 11, using the mean temperature values of the two observation levels (\bar{Y}) and the mean values of each model level (\bar{X}).

6. Pg 7. Where does eq. (14) come from?

H' is the tangent-linear version of H, as defined in Eq. 13.

7. Pg 5, 1st paragraph. There is no description of the near-surface temperature structure to introduce the reader to terms like “skin” and “subskin”.

We have rephrased and expanded the second paragraph of Sect. 4 to make the definition of skin and subskin more clear: "The different types of sensors used on satellites probe the ocean temperature at different depths. Infrared (IR) sensors measure the temperature at about 10um, a layer that is referred to as the ocean skin. Microwave (MW) sensors on the other hand measure the temperature of the layer below that, the subskin, with a depth of about 1mm. This is well below the vertical resolution of an OGCM, while these layers are strongly affected by the atmospheric conditions. [...]"

8. Pg 5, line 11. "At the same time. . .". Not always at the same time.

Rephrased: "At the same time, wind can mix ..."

9. Pg 5, line 13. "straightforward assimilation". I think you mean here that it is a problem when the observations contain significant diurnal cycle changes at the skin or subskin, and that is not accounted for when comparing the observation with the model. A straightforward approach could be to remove those observations from the assimilation as you mention later on.

By "straightforward assimilation" we mean assimilating the observations into the model without rejecting or correcting observations that are affected by the diurnal cycle.

10. Pg 6, line 8. You take the daily mean value for wind and insolation. How good is this for determining the magnitude of the diurnal cycle?

Using the mean values is of course an approximation, but one that has been used throughout the literature (e.g. Gentemann, 2003) for describing the diurnal cycle. The dependence of the magnitude of the diurnal signal on the wind and insolation categories of the CCA 00 is shown in Fig. 1.

The parameterisation of the CCA 00 bases the (sub)skin temperature estimate mostly on the shape of the temperature profile in the upper ocean layers. The categorisation in wind and insolation serves to group together similar profiles to allow for a better parameterisation. The use of the mean values is of course an approximation, but as can be seen from Fig. 1 it works reasonably well to separate the different magnitudes of the SST diurnal cycle.

11. Fig 1. What depth is the diurnal cycle that is plotted?

Added "at the subskin level".

12. Pg 6, line 11. How do you get skin and subskin estimates from GOTM?

The subskin SST represents the temperature at the base of the conductive laminar sub-layer of the ocean surface; for practical purposes we have represented this by the temperature of the top model layer of GOTM (1.5cm). The conductive sub-layer of the air-sea interface, associated with

the cool-skin effect, is parameterized and dynamically computed within GOTM to produce a modelled skin SST. Further details are provided in Pimentel et al., 2019.

13. Sec 4.3, second paragraph. Wouldn't it be fairer to do the validation of the CCA on a different time-period to the one used to build the statistics for the CCA 00?

The results in this section have been redone using profiles that were withheld from the calculation of the CCA. See also the answer to specific comment 3.

14. Section 5 is a comparison of the CCA 00 to other methods. I think its title should be changed.

The title of the section has been changed to "Performance and discussion"

15. Pg 10, line 2. Waters et al (2015) assimilate data during the day where the wind speed is high.

Indeed, Waters et al. (2015) assimilates at nighttime and also during high winds. We have clarified this in the text and included also a figure showing the comparison between SOSSTA and the upper model level method in the afternoon.

16. Last paragraph section 5. The last sentence of this paragraph describes the method that should be used in the paper to assess the performance of the CCA 00.

In the revised version of the manuscript the validation and comparison with other methods is only done using profiles that are withheld from the operator calculation.

17. Pg 11, line 24. Can you include a reference for the magnitude of the diurnal cycle?

In this location we have added a reference to Pimentel et al. 2019. The discrepancy discussed at this point is between the upper model level and the (sub)skin temperature, i.e. $\max(T_{\text{skin}} - T_{1.5\text{m}})$. As the model captures some of the variability, this is typically smaller than the magnitude of the diurnal cycle discussed in literature $T_{\text{skin,max}} - T_{\text{skin,min}}$.

Furthermore, we have added to Sect. 4 where the amplitude is discussed: "Under favourable conditions this amplitude is typically of the order of a few degrees (see e.g Flament (1994)), but values as high as 6°C have been observed (Merchant, 2008)."

Interactive comment on “Using Canonical Correlation Analysis to produce dynamically-based highly-efficient statistical observation operators” by Eric Jansen et al.

Anonymous Referee #2

Received and published: 26 February 2019

GENERAL COMMENTS

This paper essentially present a method to assimilate satellite SSTs using canonical correlation analysis (CCA); a statistical technique that permits to construct an appropriate observation operator to project the state variables of a numerical model onto observed variables. In this case the method is used to correctly assimilate measurements (satellite skin sea surface temperatures) which are not included in the set of those simulated by the model (the temperature of the first model layer) but which are physically linked to them. Based on a previous paper of Pimentel et al (2018) a 1D model (GOTM) is used to simulate high-resolution temperature profiles from which it is possible to extract temperature at the model levels (the state variable of the model) and “observed” temperatures at the sub-skin level. The skin temperature is obtained, from the sub-skin temperature, using the Fairall 1996. CCA OO is based on these simulate data rather than on real measurements.

Considering that skin temperature measurements are very rare, the idea to build a matchup dataset using a specialized model is quite interesting and represents a good compromise between the two extremes of using only very few in situ data or assimilate satellite skin or sub-skin SSTs as if the they were bulk SSTs.

The validation is based on the evaluation of ability of SOSSTA in reproducing the GOTM derived skin and sub-skin temperatures (fig. 3). SOSSTA is presented in what appears to be the companion paper of this paper still submitted to Ocean Sciences by the same authors. Probably, the author should better clarify the relation between the two papers, and rather than merging the two, indicate in some way that they are part I and II of a single subject.

The SOSSTA project is presented in three separate papers. The first (Pimentel et al. 2019) describes the modelling of the diurnal cycle in the Mediterranean Sea. The second (this paper) describes the method of building an observation operator by parameterising the results of an external model. It uses the dataset from the first paper as an example to demonstrate the method. The third paper (Korres et al., 2019) uses the GOTM datasets and the method presented in this paper and applies it in the POSEIDON data assimilation system. We believe that the topics are sufficiently different and self-contained to warrant publish them as separate papers. Moreover, we are hoping to publish additional papers such as Korres et al., 2019 that document the application of the SOSSTA operator in other models/systems.

One last, more general comment is about the fact that implications for the

diurnal cycle are well discussed in the paper also in relation with satellite data in section 4 but geo-stationary satellite, and SEVIRI in particular, are never mentioned while they should represent an interesting source of information for the assessment of the proposed assimilation procedure. However Seviri SSTs are distributed as subskin-sst. This is clearly declared in the MSG SST reprocessing ATBD v1.1, 31/5/2016 Algorithm Theoretical Basis: “Since the coefficient of the SST algorithm are established using in-situ measurements, the retrieved SST is considered to be the sub-skin SST. One could apply a -0.17C (Donlon et al., 2002) to get the skin SST. However this offset is only a very rough conversion term valid at largescale for wind speed exceeding 6 m/sec .” (osi-saf v1.1, 31/5/2016). If this sentence is correct one should verify if IR SSTs are to be considered skin or sub-skin SSTs.

Donlon, C. J., Minnett, P. J., Gentemann, C., Nightingale, T. J., Barton, I. J., Ward, B., and Murray, M. J. (2002). Toward improved and validation of satellite and sea surface and skin temperature and measurements and for climate and research. *Journal of Climate*, 15:353–359.

Following your suggestion we have included a paragraph about SEVIRI in the introduction of Sect. 4: “One of the most important sources of SST observations is the Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument onboard the Meteosat satellites of the second generation. As these are geostationary satellites, SEVIRI can provide continuous measurements of the same area with a 15-minute temporal resolution. Although the IR imager is sensitive to skin temperature, the calibration algorithm of SEVIRI corrects for the cool skin bias and the resulting SST products should be considered as subskin temperature (Saux-Picart et al., 2018). For wind speeds greater than 6 m/s the skin temperature may be calculated as $T_{\text{skin}} = T_{\text{subskin}} - 0.17^{\circ}\text{C}$ (Donlon, 2002), but this is only an approximation.”

Overall I would say that it is a good paper that deserves to be published on Ocean Sciences doing some minor revisions as suggested in this review.

SPECIFIC COMMENTS:

Page 5, section 4 Use case: satellite SST, The authors write: “Although diurnal variability is included to some extent (Marullo et al., 2014), the vertical resolution of the OGCMs is still insufficient to fully resolve the variability of the skin and subskin ocean temperature”, To resolve Skin a sub-skin is only matter of resolution or some more physics is still needed?

Vertical resolution is of key importance, but more physics is also needed. For example, the physics of the conductive skin-layer is diffusion-dominated. Pimentel et al, 2019, also explore the influence of the penetration of solar radiation within the near surface as well as the sea surface albedo. There are feedbacks in that an improved resolution of SST improves air-sea heat flux calculations.

Page 5 section 4.1: “The top 75 m of the water column is resolved using 122 vertical layers with fine resolution near the surface and gradually becoming coarser with depth. The uppermost 1 m contains a total of 21 layers, with the highest level at 1.5 cm depth”. Considering that 1.5 cm is not enough to resolve the skin and sub-skin can the author justify the choice of 122 vertical levels with the highest level at 1.5 cm depth? Is this due to computation capabilities or, for some numerical or physical reason, it makes no sense to use higher resolutions?

The subskin SST represents the temperature at the base of the conductive laminar sub-layer of the ocean surface; for practical purposes we have represented this by the temperature of the top model layer of GOTM (1.5cm). The conductive sub-layer of the air-sea interface, associated with the cool-skin effect, is parameterized and dynamically computed within GOTM to produce a modelled skin SST. Further details are provided in Pimentel et al., 2019.

There is flexibility within GOTM to explore even higher resolution, although it is not clear whether this would be justified. Additional model layers can be included, although this increases data handling and storage needs.

Page 6 section 4.2: “Under certain conditions the ocean skin may even cool down below the bulk temperature. “. “certain conditions” are related to latent heat loss.

Rephrased "Due to latent heat loss, the ocean skin may even cool down below the bulk temperature"

Page 10, section 5, lines 7-8. “This can be explained by the cool-skin effect that is included in GOTM and which plays a role also at nighttime”. Here you can cite figure 4 of Donlon et al 2002 (see reference above),

In the revised version the comparison with skin temperature has been removed, as SEVIRI provides only subskin.

Section 6, Discussion. The skill of CCA OO respect to some other method is measured using GOTM as a reference. As I already noted in the general comments this choice is, in some sense, obliged by the scarcity of in situ skin or sub-skin SST measurements. But, what about the possibility to use meteosat data as a reference?

In the revised version of the paper we are calculating the skill of the CCA OO compared to other methods using the SEVIRI L3C dataset of subskin temperature. To obtain an independent dataset we withhold every other profile (along the zonal direction) in the input dataset from the calculation of the CCA OO. The performance metrics are calculated using only the withheld profiles.

Bulk SSTs at about 20 cm of depth are routinely measured by drifters. Drifter SSTs are used to continuously assess the validity of satellite SST products,

distributed by agencies or the Copernicus Marine Service (CMEMS). Can the proposed CCA OO method also contribute to adjust drifter SSTs to skin or sub-skin temperature making more correct the comparison with the satellite estimates?

We have shown that the CCA OO method can be used to project modelled upper-ocean temperature profiles onto the skin and subskin temperature. We believe that the same approach could also be applied to temperature profiles measured by drifters.

Using Canonical Correlation Analysis to produce dynamically-based highly-efficient statistical observation operators

Eric Jansen¹, Sam Pimentel², Wang-Hung Tse², Dimitra Denaxa³, Gerasimos Korres³,
Isabelle Mirouze¹, and Andrea Storto¹

¹Euro-Mediterranean Center on Climate Change (CMCC), Italy

²Trinity Western University (TWU), Langley, BC, Canada

³Hellenic Centre for Marine Research (HCMR), Athens, Greece

Correspondence: Eric Jansen (eric.jansen@cmcc.it)

Abstract. Observation operators (OOs) are a central component of any data assimilation system. As they project the state variables of a numerical model into the space of the observations, they also provide an ideal opportunity to correct for effects that are not or not sufficiently described by the model. In such cases a dynamical OO, an OO that interfaces to a secondary and more specialised model, often provides the best results. However, given the large number of observations to be assimilated in a typical atmospheric or oceanographic model, the computational resources needed for using a fully dynamical OO mean that this option is usually not feasible. This paper presents a method, based on canonical correlation analysis (CCA), that can be used to generate highly-efficient statistical OOs that are based on a dynamical model. These OOs can provide an approximation to the dynamical model at a fraction of the computational cost.

One possible application of such an OO is the modelling of the diurnal cycle of sea surface temperature (SST) in ocean general circulation models (OGCMs). Satellites that measure SST measure the temperature of the thin uppermost layer of the ocean. This layer is strongly affected by the atmospheric conditions and its temperature can differ significantly from the water below. This causes a discrepancy between the SST measurements and the upper layer of the OGCM, which typically has a thickness of around 1 m. The CCA OO method is used to parametrise the diurnal cycle of SST. The CCA OO is based on an input dataset from the General Ocean Turbulence Model (GOTM), a high-resolution water column model that has been specifically tuned for this purpose. The parameterisations of the CCA OO are found to be in good agreement with the results from GOTM and improve upon existing parameterisations, showing the potential of this method for use in data assimilation systems.

1 Introduction

Data assimilation (DA) strives to improve the forecast skill of a numerical model by combining the model with observations. Observations are incorporated into the model by applying a series of corrections to the internal state of the model. As the state variables of a numerical model are usually not observed directly, this procedure requires an observation operator (OO) to project the model state variables onto the variable that is observed. The difference between the observation and the model prediction, the so-called innovation, forms the basis for calculating the correction to the model state. The accuracy of the OO

is paramount in this process: any bias in the projection will lead to a bias in the innovation and therefore result in a biased correction to the model state. For this reason, bias correction procedures have been built considering not only systematic errors in observations but also in observation operators (see e.g. Harris and Kelly (2001), for satellite radiance data).

Many different types of OO exist. In its simplest form, an OO could just select one of the state variables in a point near to the observation or, perhaps, perform an interpolation. More complex OOs may include corrections for processes that influence the observation, but are not or not sufficiently modelled. Ultimately one could even consider a dynamical OO that wraps a second numerical model to locally refine the results of the parent model. The latter solution may very well provide the most accurate results, but the vast number of observations that need to be assimilated in a typical atmospheric or oceanographic model means that this approach would require a prohibitive amount of computing resources. This limits OOs in most practical applications to relatively simple parameterisations in terms of the model state variables. Moreover, variational data assimilation requires observation operators to be linearised around the background within the inner loops (tangent-linear approximation). This translates into the need of building OOs that can be formally and practically differentiated.

This paper presents a method of parametrising the results of a specialised model in such a way that it can be efficiently used within an OO. The parameterisation is based on canonical correlation analysis (CCA), a well-established mathematical method for finding cross-correlations between datasets. A new pseudo-dynamical OO is generated using the canonical correlation between the inputs and outputs of the specialised model on a large and representative dataset. Once this correlation has been calculated, the application of the pseudo-dynamical OO involves only a matrix multiplication that can be performed at a fraction of the computational cost of the dynamical OO. A similar method has been used previously to build reduced-order OOs in atmospheric data assimilation (Haddad et al., 2015).

This work is part of the SOSSTA (Statistical-dynamical observation Operator for SST data Assimilation) project, funded by the EU Copernicus Marine Environment Monitoring Service (CMEMS) through the Service Evolution grants. The aim of SOSSTA is to formulate an efficient OO for SST DA that accounts for the diurnal variability of the ocean skin temperature. The project includes pilot studies in the Mediterranean Sea and the Aegean Sea.

The paper is organised as follows: Sect. 2 provides a quick review of CCA; Sect. 3 discusses how CCA can be used to construct the OO matrix; Sect. 4 applies the CCA OO to the modelling of satellite sea surface temperature (SST) measurements in oceanographic models; and Sect. 5 discusses the performance of the method and other possible applications. Conclusions are presented in Sect. 6.

2 The CCA method

CCA (Hotelling, 1936) is a method to find cross-correlations between two datasets \mathbf{X} and \mathbf{Y} . The datasets are considered to be matrices structured such that the columns represent different variables and the rows represent the measurements of these variables. CCA then aims to find transformation matrices \mathbf{A} and \mathbf{B} that transform the anomaly of the variables of \mathbf{X} and \mathbf{Y} , denoted \mathbf{X}' and \mathbf{Y}' , into the set of *canonical variables* \mathbf{F} and \mathbf{G} :

$$\mathbf{F} = \mathbf{X}'\mathbf{A} \quad \mathbf{G} = \mathbf{Y}'\mathbf{B} \tag{1}$$

The structure of \mathbf{F} and \mathbf{G} matches that of \mathbf{X} and \mathbf{Y} . The canonical variables are constructed such that the variable \mathbf{F}_i is maximally correlated to the variable \mathbf{G}_i . At the same time, both \mathbf{F}_i and \mathbf{G}_i are uncorrelated to \mathbf{F}_j and \mathbf{G}_j for $i \neq j$; therefore each additional canonical variable describes the maximal remaining correlation between the two datasets. The number of canonical variables that can be obtained with this procedure is limited to the smallest number of variables in \mathbf{X} or \mathbf{Y} .

- 5 The calculation of the matrices \mathbf{A} and \mathbf{B} is relatively straightforward using the algorithm of Björck and Golub (1973). Writing the requirements outlined above in equation form yields:

$$\mathbf{F}^T \mathbf{F} = \mathbf{G}^T \mathbf{G} = \mathbf{I} \quad (2a)$$

$$\mathbf{F}^T \mathbf{G} = \mathbf{D} \quad (2b)$$

- with \mathbf{I} the unit matrix and \mathbf{D} a diagonal matrix. The algorithm uses a QR-decomposition to decompose both \mathbf{X}' and \mathbf{Y}' into an orthogonal matrix \mathbf{Q} and an upper-triangular matrix \mathbf{R} :

$$\mathbf{X}' = \mathbf{Q}_x \mathbf{R}_x \quad \mathbf{Y}' = \mathbf{Q}_y \mathbf{R}_y \quad (3)$$

The algorithm proceeds by applying a Singular Value Decomposition (SVD) on the product $\mathbf{Q}_x^T \mathbf{Q}_y$:

$$\mathbf{Q}_x^T \mathbf{Q}_y = \mathbf{U} \mathbf{S} \mathbf{V}^T \quad (4)$$

Trying the Ansatz:

$$15 \quad \mathbf{A} \equiv \mathbf{R}_x^{-1} \mathbf{U} \quad \mathbf{B} \equiv \mathbf{R}_y^{-1} \mathbf{V} \quad (5)$$

the orthonormality requirement of Eq. 2a becomes:

$$\begin{aligned} \mathbf{F}^T \mathbf{F} &= \mathbf{A}^T \mathbf{X}'^T \mathbf{X}' \mathbf{A} \\ &= \left(\mathbf{U}^T (\mathbf{R}_x^{-1})^T \right) (\mathbf{R}_x^T \mathbf{Q}_x^T) (\mathbf{Q}_x \mathbf{R}_x) (\mathbf{R}_x^{-1} \mathbf{U}) \\ &= \mathbf{I} \end{aligned} \quad (6)$$

- 20 and an analogous result follows for $\mathbf{G}^T \mathbf{G}$. The orthogonality requirement of Eq. 2b becomes:

$$\begin{aligned} \mathbf{D} = \mathbf{F}^T \mathbf{G} &= \mathbf{A}^T \mathbf{X}'^T \mathbf{Y}' \mathbf{B} \\ &= \left(\mathbf{U}^T (\mathbf{R}_x^{-1})^T \right) (\mathbf{R}_x^T \mathbf{Q}_x^T) (\mathbf{Q}_y \mathbf{R}_y) (\mathbf{R}_y^{-1} \mathbf{V}) \\ &= \mathbf{U}^T \left(\mathbf{U} \mathbf{S} \mathbf{V}^T \right) \mathbf{V} = \mathbf{S} \end{aligned} \quad (7)$$

- Therefore the Ansatz of Eq. 5 is a valid solution for the matrices \mathbf{A} and \mathbf{B} . Moreover, by counting the number of degrees of freedom in these matrices and the number of constraints provided by Eq. 2, it can be shown that all solutions are permutations of Eq. 5 (Press, 2011). The canonical basis is therefore uniquely defined. In case that \mathbf{X} and \mathbf{Y} contain different numbers of variables N_x and N_y , the SVD of Eq. 4 selects the N largest correlations, with $N = \min(N_x, N_y)$.

As QR-decomposition and SVD are common matrix operations that are efficiently implemented in most numerical libraries, this algorithm is straightforward to implement in most programming languages.

3 Using CCA to construct an OO

The CCA method can be used to construct an OO. Let \mathbf{X} be a set of (possibly) relevant model state variables and \mathbf{Y} the corresponding observation values. Here \mathbf{Y} could be obtained from a specialised model, but also from a historical dataset of real observations. Applying the algorithm of Sec. 2 yields the matrices \mathbf{A} , \mathbf{B} and \mathbf{D} . The first two convert the mean-subtracted model states \mathbf{X}' and observation values \mathbf{Y}' into their canonical counterparts \mathbf{F} and \mathbf{G} . The diagonal matrix \mathbf{D} holds for each pair of canonical variables i the best fit to the slope of the correlation: $\mathbf{D}_{ii} = d\mathbf{G}_i/d\mathbf{F}_i$.

Assuming that $N_x \geq N_y$ —i.e. the number of model state variables is at least equal to the number of observed variables—it is possible to calculate \mathbf{Y}' from \mathbf{X}' by passing through canonical space and applying the fitted slope \mathbf{D} :

$$\mathbf{Y}' = \mathbf{X}' \mathbf{A} \mathbf{D} \mathbf{B}^{-1} \equiv \mathbf{X}' \mathbf{M}, \quad (8)$$

10 defining the CCA OO matrix:

$$\mathbf{M} \equiv \mathbf{A} \mathbf{D} \mathbf{B}^{-1}, \quad (9)$$

of size $N_x \times N_y$. As the CCA considers only the anomaly of \mathbf{X} and \mathbf{Y} , an additional offset term needs to be considered to accommodate the mean values of \mathbf{X} and \mathbf{Y} in the input dataset. However, the mean values of \mathbf{X} and \mathbf{Y} can be combined by applying the matrix \mathbf{M} :

$$\begin{aligned} 15 \quad \mathbf{Y} - \bar{\mathbf{Y}} &= (\mathbf{X} - \bar{\mathbf{X}}) \mathbf{M} \\ \mathbf{Y} &= \mathbf{X} \mathbf{M} + \mathbf{K} \end{aligned} \quad (10)$$

with:

$$\mathbf{K} \equiv \bar{\mathbf{Y}} - \bar{\mathbf{X}} \mathbf{M} \quad (11)$$

a combined offset vector of length N_y .

20 During the training-phase of the CCA OO, the datasets \mathbf{X} and \mathbf{Y} are used to calculate the matrix \mathbf{M} and the offset \mathbf{K} . Once computed, they can be used to form an observation operator \mathbf{H} that transforms a state \mathbf{x} as:

$$\mathbf{H}(\mathbf{x}) = \mathbf{x} \mathbf{M} + \mathbf{K} \quad (12)$$

Furthermore, the tangent-linear approximation used in variational DA schemes requires that:

$$\mathbf{H}(\mathbf{x}) \sim \mathbf{H}(\mathbf{x}^b) + \mathbf{H}' d\mathbf{x} \quad (13)$$

25 where \mathbf{H}' is the tangent-linear version of the OO, \mathbf{x}^b the background state and $d\mathbf{x}$ the deviation from the background. The CCA OO is straightforward to implement in this scheme, since [for \$\mathbf{H}'\$ and its adjoint \$\mathbf{H}'^T\$ it follows that:](#)

$$\mathbf{H}' = \mathbf{M}^T \quad \mathbf{H}'^T = \mathbf{M} \quad (14)$$

4 Use case: satellite SST

One possible application of the new CCA OO is the assimilation of SST in Ocean General Circulation Models (OGCMs). In recent years OGCMs have seen significant improvements in vertical resolution, particularly near the surface, where the first layer has been reduced to a thickness of the order of 1 m or less. At this resolution, the diurnal cycle of SST should be taken into account. Although diurnal variability is included to some extent (Marullo et al., 2014), the vertical resolution of the OGCMs is still insufficient to fully resolve the variability of the skin and subskin ocean temperature.

This issue becomes particularly evident when assimilating satellite SST observations. The different types of sensors used on satellites probe the ocean temperature at different depths. Infrared (IR) sensors measure the temperature at about 10 μm , a layer that is referred to as the ocean skin. Microwave (MW) sensors on the other hand measure the temperature of the layer below that, the subskin, with a depth of about 1 mm. This is much shallower than the vertical resolution of a typical OGCM, while these layers are strongly affected by the atmospheric conditions. The ocean skin cools due to thermodynamic processes at the air-sea interface, while the absorption of solar heat causes a warming of the subskin. At the same time, wind can mix the skin and subskin with the water below, smoothing the temperature variations. During days of low wind and/or high insolation conditions the amplitude of the SST diurnal cycle can be larger than the combined accuracy of the model and observations, causing a straightforward assimilation of SST to degrade the performance of the model (Marullo et al., 2016). Under favourable conditions this amplitude is typically of the order of a few degrees (see e.g. Flament et al. (1994)), but values as high as 6° C have been observed (Merchant et al., 2008).

Errors due to e.g. limited spatial resolution or unrepresented processes are generally included in the representation error. Representation errors have been extensively discussed within ocean applications (Oke and Sakov, 2008; Janjić et al., 2018). However, the diurnal variability of skin SST represents a potentially systematic error that requires a proper treatment rather than just the increase the representation component of the observational error.

An important source of SST observational data is the Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument onboard the Meteosat satellites of the second generation. As these are geostationary satellites, SEVIRI can provide continuous measurements of the same area with a 15-minute temporal resolution. Although the IR imager is sensitive to skin temperature, the calibration algorithm of SEVIRI corrects for the cool skin bias and the resulting SST products should be considered as subskin temperature (Saux Picart and Legendre, 2018). For wind speeds greater than 6 m/s the skin temperature may be calculated as $T_{\text{skin}} = T_{\text{subskin}} - 0.17^\circ \text{C}$ (Donlon et al., 2002), but this is only an approximation.

This section will discuss how to use the output of a water column model specifically tuned for modelling the diurnal cycle of SST together with the CCA OO to build an observation operator for SST that accounts for the diurnal variability.

30 4.1 General Ocean Turbulence Model

The SST diurnal cycle is modelled using the General Ocean Turbulence Model (GOTM). GOTM is a one-dimensional water column model that includes multiple turbulence closure schemes (Burchard et al., 1999; Umlauf et al., 2005). It has been successfully adapted to model the near-surface variability of ocean temperature, including both the diurnal cycle and the cool-skin

effect (Pimentel et al., 2008a, b). Recently it has been used to systematically simulate the atmospheric and oceanographic conditions in the Mediterranean Sea (Pimentel et al., 2019). The latter study has resulted in a multi-year dataset, modelling the diurnal cycle in the Mediterranean Sea on a grid of $0.75^\circ \times 0.75^\circ$ resolution and with hourly time resolution. For this dataset GOTM is configured with the $k-\varepsilon$ turbulent kinetic energy parameterisation with internal waves. The top 75 m of the water column is resolved using 122 vertical layers with fine resolution near the surface and gradually becoming coarser with depth. The uppermost 1 m contains a total of 21 layers, with the highest level at 1.5 cm depth. This dataset is used in the present paper to build the CCA OO for SST.

The subskin SST represents the temperature at the base of the conductive laminar sub-layer of the ocean surface; for practical purposes it is represented by the temperature of the top model layer of GOTM (1.5 cm). The conductive sub-layer of the air-sea interface, associated with the cool-skin effect, is parameterised and dynamically computed within GOTM to produce a modelled skin SST. Further details are provided in Pimentel et al. (2019).

4.2 Operator setup

The aim for the CCA OO is to parameterise the IR and MW satellite SST observations as a function of temperature in the water column below. While the dataset of Pimentel et al. (2019) uses a fine vertical resolution to calculate the SST observations, the CCA OO will consider only the levels of a typical OGCM. Within the SOSSTA project this OGCM is the CMEMS Mediterranean Forecasting System (MFS) (Simoncelli et al., 2014), but the parameterisation can be performed for any vertical distribution of levels.

The magnitude of the diurnal signal depends strongly on the atmospheric conditions, most importantly the insolation and wind speed. Insolation causes the ocean skin to heat up during the course of the day, while wind mixes the upper layers of the ocean leading to dissipation of the heat. Due to latent heat loss, the ocean skin may even cool down below the bulk temperature. To accommodate non-linear dependence on the different insolation and wind scenarios in the CCA OO, the GOTM dataset is divided into 12 insolation and 8 wind categories. Insolation and wind are defined in each location as the daily mean value in local mean time (LMT). The category boundaries were chosen to equally divide the dataset. The magnitude of the diurnal warming for the different categories is shown in Fig. 1.

The GOTM dataset has been compared to SEVIRI data at the skin level in Pimentel et al. (2019) and was found to be in good agreement over the whole period of 2013 and 2014. However, after dividing the dataset in atmospheric categories it is found that categories with high diurnal warming may have a warm bias of up to 0.5°C and categories with low diurnal warming a cold bias of typically $0.1\text{--}0.2^\circ\text{C}$. This category bias is corrected for by subtracting the mean difference between SEVIRI and GOTM at subskin level for each category.

For each category of wind and insolation, and at hourly time resolution, the CCA OO is calculated to project the 10 uppermost levels of the MFS model onto the skin and subskin SST temperatures. The 10 levels extend down to a depth of approximately 40 m, which was chosen to be well below the depth influenced by the diurnal cycle of temperature. Figure 2(a) shows the correlation between the model temperature at various depths and the two SST observation types. As expected, the SST is strongly correlated to the highest levels and the correlation decreases with depth. It is important to note that in this case the

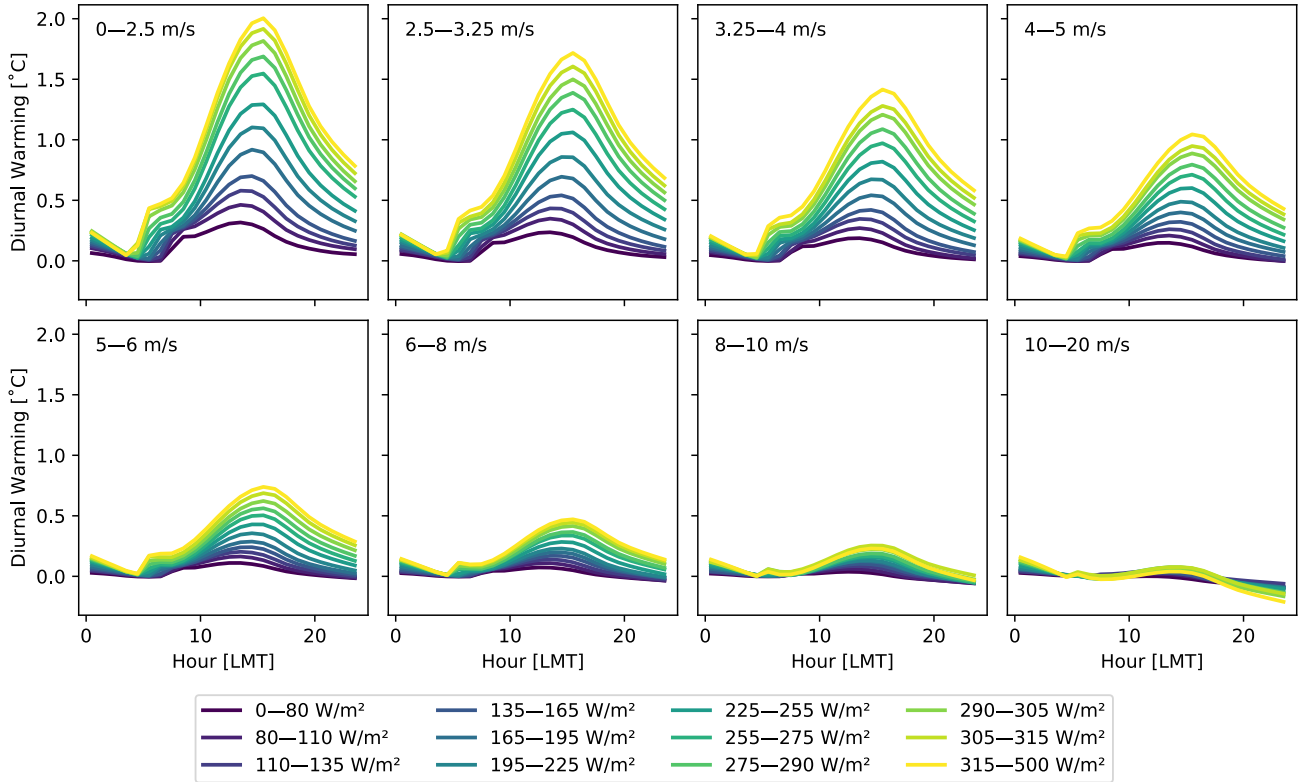


Figure 1. The magnitude of the diurnal warming [at the subskin level](#) as a function of the time of the day for different wind and insolation categories. The diurnal warming is measured with respect to the SST at local sunrise. The wind categories are represented by the different panels, while the insolation categories are shown as different curves within each panel.

various levels are also strongly correlated to each other. Figure 2(b) shows the correlation after transforming to canonical coordinates. It can be seen that the strongest correlation has not significantly changed, as the first canonical variable is very similar to the highest model level. The second pair of canonical variables ($\mathbf{F}_2, \mathbf{G}_2$), however, describes an additional correlation of around 60% between model water temperature and SST.

5 4.3 Validation

The CCA OO is validated by comparing its performance to that of the full GOTM. To be able to use the operator effectively in a DA system, it should be able to provide an accurate approximation of the GOTM results. [The validation is performed against GOTM profiles that are withheld from the CCA OO calculation. The GOTM dataset is split in two, withholding every other profile in the zonal direction from the calculation. The validation then uses the withheld profiles](#) and extracts the 10 depths corresponding to the MFS levels, mimicking the use of the operator inside a DA system. The CCA OO, based on the

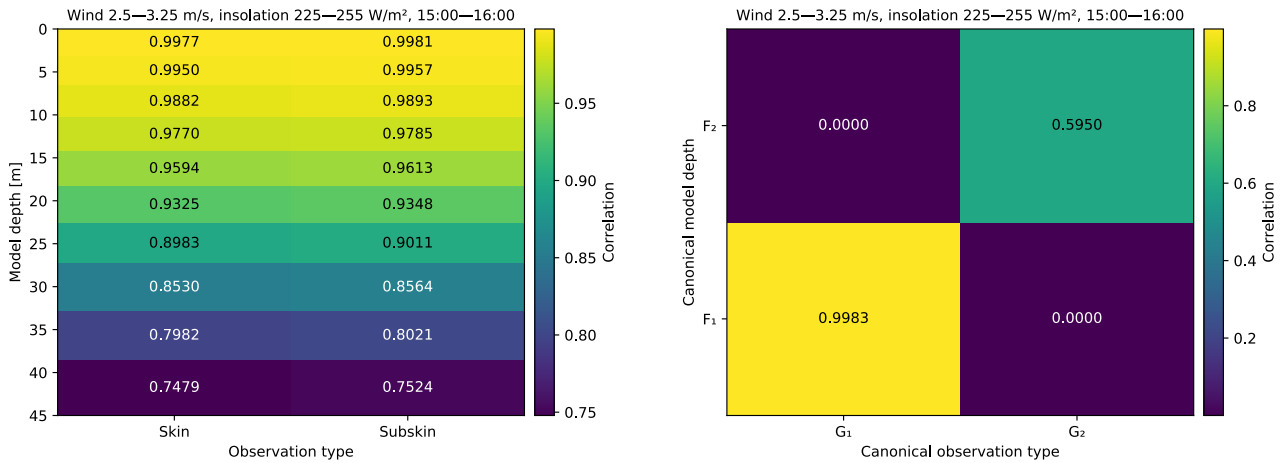


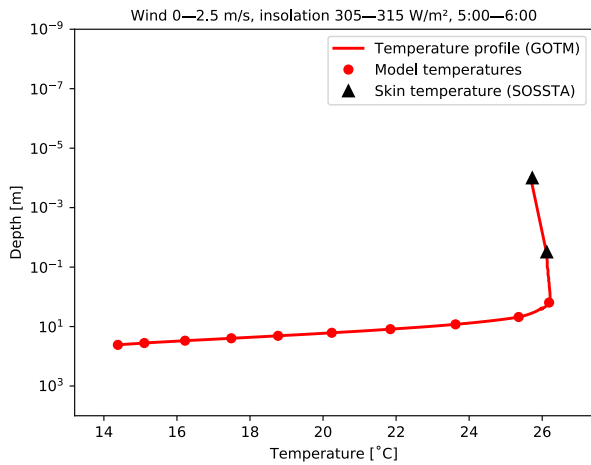
Figure 2. The correlation coefficients between the model variables and observations (left); and the canonical equivalent of these variables (right).

atmospheric category and closest time, is subsequently applied to project the model temperature onto the skin and subskin SST. The projected SST values are then compared to the values in the original GOTM profile.

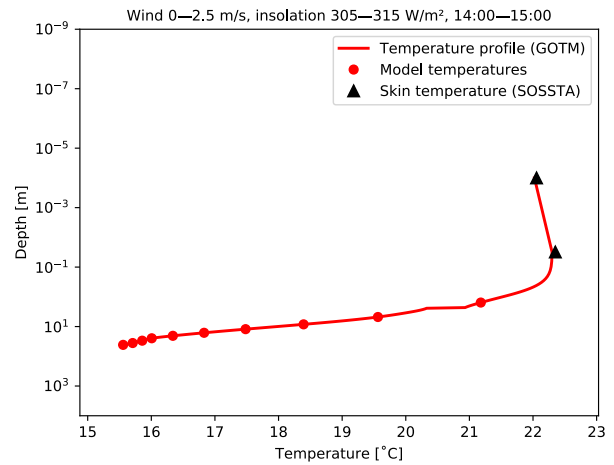
Some examples of the validation are shown in Fig. 3. Each panel shows a profile from the GOTM dataset, together with the model levels that were used as input to the CCA OO. The output of the CCA OO is superimposed onto the GOTM profile, so that a comparison can be made. Figure 3(a) shows a temperature profile in the early morning, during a day of low wind and high insolation. At this time the diurnal warming is limited and due to the clear sky conditions the skin and subskin temperatures have cooled down slightly below the temperature of the first model level. Figure 3(b) shows an afternoon profile on a similar day. At this time the diurnal warming is around its maximum and the skin temperature has increased about 1° C above the first level of the model. In case of high wind speed, the increased mixing of the upper layer of the ocean can completely cancel the effect of the high insolation, as is shown in Fig. 3(c). In this situation the temperature in the upper 10 m of the ocean is almost constant. When high wind conditions coincide with low insolation, the surface can also cool quite significantly, as is shown in Fig. 3(d). The CCA OO is able to reproduce correctly the GOTM skin and subskin temperature under different atmospheric conditions. The atmospheric categories with strong diurnal warming have a Root Mean Square Error (RMSE) of up to 0.4°C, for all other categories the RMSE is around 0.1°C. The bias of the CCA OO compared to GOTM was found to be negligible.

5 Performance and discussion

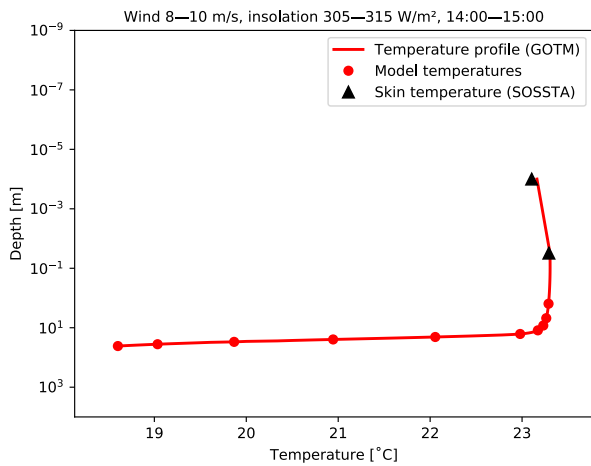
The performance of the GOTM-based CCA OO for SST is compared to other commonly used methods. [For this comparison the GOTM dataset is again split along the zonal direction, using every other profile to calculate the CCA OO. The remaining profiles are matched to SEVIRI subskin retrievals, using only profiles matched to a measurement with acceptable \(4\) or](#)



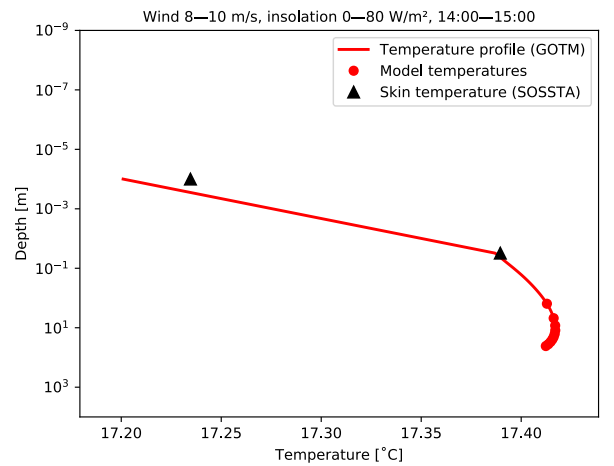
(a) Low wind, high insolation, early morning



(b) Low wind, high insolation, afternoon



(c) High wind, high insolation, afternoon



(d) High wind, low insolation, afternoon

Figure 3. Examples of temperature profiles in various conditions and at different times. The GOTM profiles are shown by the red curve, while the filled circles indicate the values used as input to the CCA OO. The output of the CCA OO is shown by the black triangles.

[good \(5\) quality control level](#). The performance can be conveniently expressed in terms of the skill score (SS), defined by Murphy (1988) as:

$$SS = 1 - \frac{MSE_{\text{model}}}{MSE_{\text{reference}}} \quad (15)$$

The skill score is based on the mean square error (MSE) of the model under test and of a reference model. Specifically, it expresses the difference in MSE as a fraction of the reference MSE. The skill score is straightforward to interpret: a perfect model ($MSE = 0$) results in a skill score of 1, while a model that shows no improvement over the reference model receives a skill score of 0. Negative skill scores indicate that the model performs worse and its MSE has increased with respect to the reference.

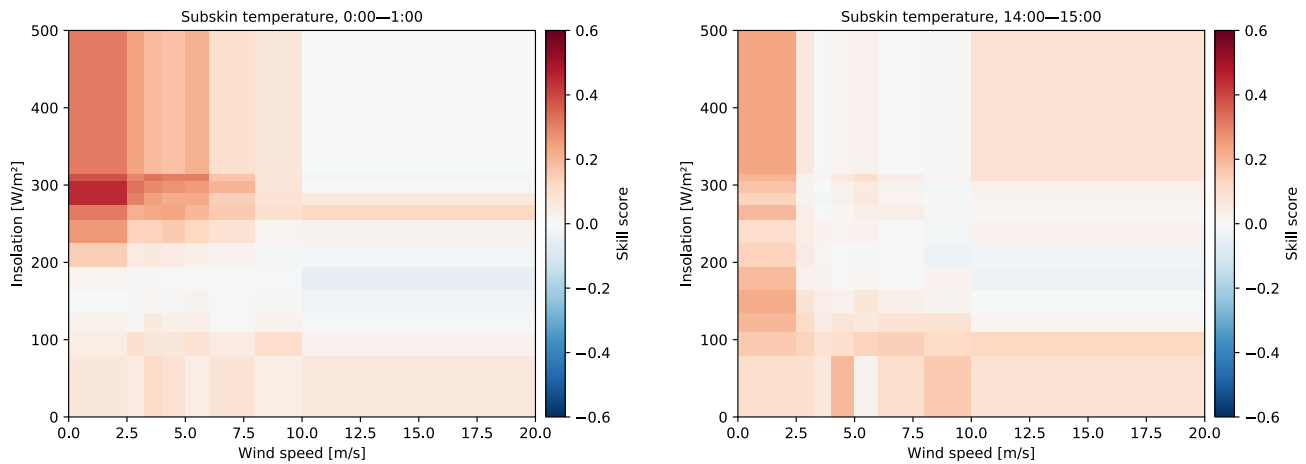


Figure 4. Skill score of the CCA OO compared to upper model level for all wind and insolation categories at midnight (left) and in the afternoon (right).

The simplest method of assimilating satellite SST observations in a model that insufficiently describes the diurnal cycle of SST is to assimilate only at night [or during high wind](#), see for example Waters et al. (2015). During the night the cycle of SST is close to its minimum value and the temperature of the upper model layer forms a reasonable approximation for the skin temperature. In this situation the assimilation is performed without additional corrections. Figure 4(a) shows the skill score of the CCA OO at midnight local time, using as reference method the upper model layer. For both methods the MSE is calculated with respect to the [SEVIRI subskin](#) temperature. [Figure 4\(b\) shows the same situation, but in the afternoon. For high wind and low insolation the CCA OO performs, as expected, similarly to using the upper model layer. However, for low wind speeds and high insolation the CCA OO shows a clear improvement, even at midnight. This can be explained by the fact that at midnight still some diurnal signal remains and, even using the wind and insolation values of the next day, this is correctly modelled by the CCA OO.](#)

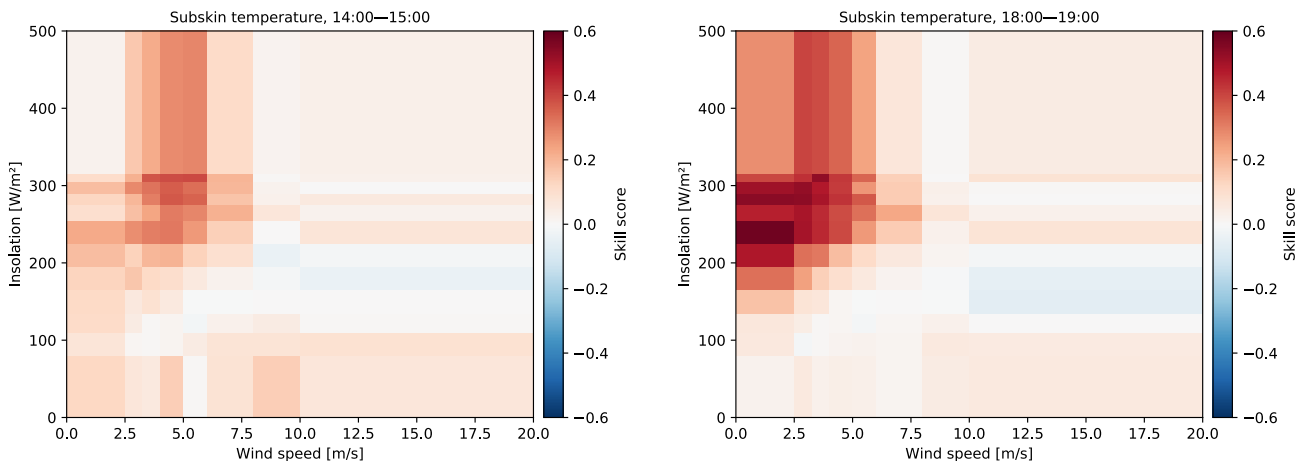


Figure 5. Skill score of the CCA OO compared to the parameterisation of Bernie et al. (2007).

A more advanced solution is the parameterisation of Bernie et al. (2007), which estimates the diurnal signal as a function of wind, insolation and time. This is a commonly used parameterisation, for example included with the NEMO ocean model (Madec et al., 1998). Figure 5 shows the skill score for the CCA OO compared to the parameterisation of Bernie et al. (2007) at the peak of the diurnal cycle (a) and in the early evening (b). It can be seen that for high insolation and low wind, conditions for which the diurnal warming is largest, both methods perform similarly. However, the CCA OO is better at accommodating different atmospheric conditions and shows significant improvements for the intermediate insolation and wind categories. Moreover, Fig. 5(b) shows that the CCA OO is able to better parametrise the cooling of the subskin in the late afternoon/evening after the peak of the diurnal warming has passed.

Using the CCA OO to improve the description of SST has many potential applications. For example, the CCA OO could be used as a parameterisation of diurnally varying skin SST within an OGCM as part of the air-sea flux calculations. The skin SST is the true interface temperature for air-sea fluxes, so this approach should result in improved air-sea heat transfer in OGCMs and coupled ocean-atmosphere models. See for example Marullo et al. (2016). Another possibility would be the use of the CCA OO as a parameterisation of diurnally varying SST within a climate model. The diurnal cycle is a fundamental signal of the climate system, yet for climate models the lack of vertical structure (and temporal resolution) is even more critical. See for example Large and Caron (2015).

Due to the way in which it is constructed, the CCA OO is an inherently linear operator. This makes it straightforward to implement in DA schemes that require linearised and differentiable OOs. However, non-linear effects can be accommodated to some extent by constructing a series of CCA OOs conditioned on such a non-linear dependency. For example, in the case of SST, this method has been used to condition the CCA OO on insolation, wind and time. The only requirement in this case is that the datasets \mathbf{X} and \mathbf{Y} of Sec. 3 are sufficiently large to divide them by such a dependent variable.

The minimum size of the input dataset required depends ultimately on the number of model variables used (N_x) and the number of observation variables to predict (N_y). The number of free parameters in the CCA OO matrix \mathbf{M} and the offset \mathbf{K} equals $(N_x + 1)N_y$. As each entry in the input dataset also provides N_y observation values, Eq. 4 requires a minimum of $N_x + 1$ entries to be mathematically solvable. However, at this point the CCA OO will be overfitted. It will simply be able to memorise the input datasets rather than being based on general characteristics of the data. Care has to be taken to avoid this situation, making sure the input dataset contains a number of entries n with $n \gg N_x$. Whether a given size n is sufficient should be tested using independent data. One possible method for this test is to withhold part of the input dataset from the CCA OO calculation, then use this subset to calculate the CCA OO performance.

6 Conclusions

Observation operators (OOs) form a central component in any data assimilation (DA) system, as they transform the state variables of a numerical model into real-world observable variables. Often an OO also needs to correct for processes that are not fully described by the parent model. Such processes may be best modelled by interfacing the OO to a specialised model, but this is generally not feasible due to computational constraints.

The assimilation of satellite Sea Surface Temperature (SST) in ocean general circulation models (OGCMs) is a prime example of a situation where insufficiently modelled processes play an important role. The diurnal cycle of SST causes a discrepancy in the temperature of the very thin upper layer measured by the satellite and the rather coarse upper layer in a typical OGCM. On a clear summer day with low wind, this discrepancy can amount to as much as 2° C or more ([Pimentel et al., 2019](#)).

The current paper presented a method, based on Canonical Correlation Analysis (CCA), to build parameterisations based on an output dataset of a specialised model. These parameterisations, referred to as the CCA OO, can provide an efficient approximation to the results of the specialised model and are therefore well-suited for use in DA systems.

The case of SST assimilation has been used to demonstrate the new CCA OO. Using an output dataset of the General Ocean Turbulence Model (GOTM), a high-resolution water column model specifically tuned for modelling the diurnal cycle of SST, a new CCA OO has been derived. Subsequently, the operator has been applied to reduced-resolution temperature profiles from GOTM to simulate its use in a DA system. The approximations provided by the CCA OO are found to be in good agreement with the GOTM model at various times of the day and across all atmospheric conditions. The results indicate that the CCA OO could be used to enable the assimilation of SST under conditions where this was previously not possible. Moreover, the atmospheric categories that were introduced in the construction of the CCA OO for SST show that the linear assumption implicit in CCA can be partially relaxed. This makes the CCA OO versatile for any condition. Compared to commonly used methods for SST assimilation, the CCA OO can provide substantial improvements. This is especially true for measurements of the skin SST, since the CCA OO profits from the modelling of the cool-skin effect that is included in GOTM.

The ability of the CCA OO to handle complicated physical models in a relatively simple way is attractive for a large number of problems in DA, where reduced-order OOs are desirable due to computational constraints. Remotely sensed data are the obvious target, given the complexity of their relationships with state variables. Observations in coupled assimilation (e.g.

ocean-atmosphere, ocean-sea-ice or ocean-biogeochemistry) are examples of challenging problems that could be investigated in the future with the CCA OO.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work forms part of the SOSSTA project which has been funded by the EU Copernicus Marine Environment Monitoring Service (CMEMS) through the Service Evolution grants.

References

- Bernie, D. J., Guilyardi, E., Madec, G., Slingo, J. M., and Woolnough, S. J.: Impact of resolving the diurnal cycle in an ocean–atmosphere GCM. Part 1: a diurnally forced OGCM, *Climate Dynamics*, 29, 575–590, <https://doi.org/10.1007/s00382-007-0249-6>, 2007.
- Björck, Å. and Golub, G. H.: Numerical Methods for Computing Angles Between Linear Subspaces, *Mathematics of Computation*, 27, 579–594, <https://doi.org/10.2307/2005662>, 1973.
- Burchard, H., Bolding, K., and Ruiz-Villarreal, M.: GOTM, a general ocean turbulence model. Theory, implementation and test cases., Tech. Rep. EUR 18745 EN, European Commission, Brussels, Belgium, 1999.
- Donlon, C. J., Minnett, P. J., Gentemann, C., Nightingale, T. J., Barton, I. J., Ward, B., and Murray, M. J.: Toward Improved Validation of Satellite Sea Surface Skin Temperature Measurements for Climate Research, *Journal of Climate*, 15, 353–369, [https://doi.org/10.1175/1520-0442\(2002\)015<0353:TIVOSS>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<0353:TIVOSS>2.0.CO;2), [http://dx.doi.org/10.1175/1520-0442\(2002\)015<0353:TIVOSS>2.0.CO;2](http://dx.doi.org/10.1175/1520-0442(2002)015<0353:TIVOSS>2.0.CO;2), 2002.
- Flament, P., Firing, J., Sawyer, M., and Trefois, C.: Amplitude and Horizontal Structure of a Large Diurnal Sea Surface Warming Event during the Coastal Ocean Dynamics Experiment, *Journal of Physical Oceanography*, 24, 124–139, [https://doi.org/10.1175/1520-0485\(1994\)024<0124:AAHSA>2.0.CO;2](https://doi.org/10.1175/1520-0485(1994)024<0124:AAHSA>2.0.CO;2), <https://journals.ametsoc.org/doi/abs/10.1175/1520-0485%281994%29024%3C0124%3AAHSA%3E2.0.CO%3B2>, 1994.
- Haddad, Z. S., Steward, J. L., Tseng, H. C., Vukicevic, T., Chen, S. H., and Hristova-Veleva, S.: A data assimilation technique to account for the nonlinear dependence of scattering microwave observations of precipitation, *Journal of Geophysical Research: Atmospheres*, 120, 5548–5563, <https://doi.org/10.1002/2015JD023107>, 2015.
- Harris, B. A. and Kelly, G.: A satellite radiance-bias correction scheme for data assimilation, *Quarterly Journal of the Royal Meteorological Society*, 127, 1453–1468, <https://doi.org/10.1002/qj.49712757418>, 2001.
- Hotelling, H.: Relations Between Two Sets of Variates, *Biometrika*, 28, 321–377, 1936.
- Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, *Quarterly Journal of the Royal Meteorological Society*, 144, 1257–1278, <https://doi.org/doi:10.1002/qj.3130>, <https://doi.org/10.1002/qj.3130>, 2018.
- Korres, G., Denaxa, D., Jansen, E., Mirouze, I., Pimentel, S., Tse, W.-H., and Storto, A.: Assimilation of SST data in the POSEIDON system using the SOSSTA statistical-dynamical observation operator, *Ocean Science*, Submitted, 2018.
- Large, W. G. and Caron, J. M.: Diurnal cycling of sea surface temperature, salinity, and current in the CESM coupled climate model, *Journal of Geophysical Research: Oceans*, 120, 3711–3729, <https://doi.org/10.1002/2014JC010691>, 2015.
- Madec, G., Delecluse, P., Imbard, M., and Lévy, C.: OPA 8.1 Ocean General Circulation Model Reference Model, Tech. Rep. 11, Institut Pierre Simon Laplace des Sciences de l’Environnement Global, 1998.
- Marullo, S., Santoleri, R., Ciani, D., Borgne, P. L., Péré, S., Pinardi, N., Tonani, M., and Nardone, G.: Combining model and geostationary satellite data to reconstruct hourly SST field over the Mediterranean Sea, *Remote Sensing of Environment*, 146, 11–23, <https://doi.org/https://doi.org/10.1016/j.rse.2013.11.001>, 2014.
- Marullo, S., Minnett, P. J., Santoleri, R., and Tonani, M.: The diurnal cycle of sea-surface temperature and estimation of the heat budget of the Mediterranean Sea, *Journal of Geophysical Research: Oceans*, 121, 8351–8367, <https://doi.org/10.1002/2016JC012192>, 2016.
- Merchant, C. J., Filippiak, M. J., Le Borgne, P., Roquet, H., Autret, E., Piollé, J. F., and Lavender, S.: Diurnal warm-layer events in the western Mediterranean and European shelf seas, *Geophysical Research Letters*, 35, <https://doi.org/doi:10.1029/2007GL033071>, 2008.

- Murphy, A. H.: Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient, *Monthly Weather Review*, 116, 2417–2424, [https://doi.org/10.1175/1520-0493\(1988\)116<2417:SSBOTM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2), 1988.
- Oke, P. R. and Sakov, P.: Representation Error of Oceanic Observations for Data Assimilation, *Journal of Atmospheric and Oceanic Technology*, 25, 1004–1017, <https://doi.org/10.1175/2007JTECHO558.1>, <https://doi.org/10.1175/2007JTECHO558.1>, 2008.
- 5 Pimentel, S., Haines, K., and Nichols, N. K.: Modeling the diurnal variability of sea surface temperatures, *Journal of Geophysical Research: Oceans*, 113, C11 004, <https://doi.org/10.1029/2007JC004607>, 2008a.
- Pimentel, S., Haines, K., and Nichols, N. K.: The assimilation of satellite-derived sea surface temperatures into a diurnal cycle model, *Journal of Geophysical Research: Oceans*, 113, C09 013, <https://doi.org/10.1029/2007JC004608>, 2008b.
- Pimentel, S., Tse, W.-H., Xu, H., Denaxa, D., Jansen, E., Korres, G., Mirouze, I., and Storto, A.: Modeling the near-surface
10 diurnal cycle of sea surface temperature in the Mediterranean Sea, *Journal of Geophysical Research: Oceans*, 124, 171–183, <https://doi.org/10.1029/2018JC014289>, 2019.
- Press, W. H.: Canonical Correlation Clarified by Singular Value Decomposition, <http://numerical.recipes/whp/workingpapers.html>, 2011.
- Saux Picart, S. and Legendre, G.: MSG/SEVIRI Sea Surface Temperature data record Product User Manual, Tech. Rep. OSI-250, EUMETSAT, OSI SAF, https://doi.org/10.15770/EUM_SAF_OSI_0004, 2018.
- 15 Simoncelli, S., Fratianni, C., Pinardi, N., Grandi, A., Drudi, M., Oddo, P., and Dobricic, S.: Mediterranean Sea physical reanalysis (MEDREA 1987-2015) (Version 1), Tech. rep., EU Copernicus Marine Service Information, https://doi.org/10.25423/medsea_reanalysis_phys_006_004, 2014.
- Umlauf, L., Burchard, H., and Bolding, K.: General Ocean Turbulence Model, Scientific Documentation v3.2., Tech. Rep. 63, Institute for Baltic Sea Research Warnemünde, Rostock-Warnemünde, Germany, 2005.
- 20 Waters, J., Lea, D. J., Martin, M. J., Mirouze, I., Weaver, A., and While, J.: Implementing a variational data assimilation system in an operational 1/4 degree global ocean model, *Quarterly Journal of the Royal Meteorological Society*, 141, 333–349, <https://doi.org/10.1002/qj.2388>, 2015.