Online Learning of Wind-Field Models

Lehel Csató, Dan Cornford, and Manfred Opper

Neural Computing Research Group, Aston University
B4 7ET Birmingham, United Kingdom
{csatol,cornfosd,opperm}@aston.ac.uk

Abstract. We study online approximations to Gaussian process models for spatially distributed systems. We apply our method to the prediction of wind fields over the ocean surface from scatterometer data. Our approach combines a sequential update of a Gaussian approximation to the posterior with a sparse representation that allows to treat problems with a large number of observations.

1 Introduction

A common scenario of applying online or sequential learning methods [Saad 1998] is when the amount of data is too large to be processed by more efficient offline methods or there is no possibility to store the arriving data. In this article we consider the area of spatial statistics [Cressie 1991], where the data is observed at different spatial locations and the aim is to build a global Bayesian model of the local observations based on a Gaussian Process prior distribution. Specifically, we consider scatterometer data obtained from the ERS-2 satellite [Offiler 1994] where the aim is to obtain an estimate of the wind fields which the scatterometer indirectly measured.

The scatterometer measures the radar backscatter from the ocean surface at a wavelength of approximately 5 $\mathtt{cm}$. The strength of the returned signal gives an indication of the wind speed and direction, relative to scatterometer beam direction. As shown in [Stoffelen and Anderson 1997b] the measured backscatter behaves as a truncated Fourier expansion in relative wind direction. Thus while the wind vector to scatterometer observations map is one-to-one, its inverse is one-to-many [Evans *et al.* 2000]. This makes the retrieval of a wind field a complex problem with multiple solutions. Nabney *et al.* [2000] have recently proposed a Bayesian framework for wind field retrieval combining a vector Gaussian process prior model with local forward (wind field to scatterometer) or inverse models.

One problem with the approach outlined in [Nabney *et al.* 2000] is that the vector Gaussian process requires a matrix inversion which scales as $\mathtt{n}^3$. The backscatter is measured over 50×50 $\mathtt{km}$ cells over the ocean and the total number of observations acquired on a given orbit can be several thousand.

In this paper we show that we can produce an efficient approximation to the posterior distribution of the wind field by applying a Bayesian online learning approach [Opper 1998] to Gaussian process models following [Csató and Opper

2001], which computes the approximate posterior by a single sweep through the data. The computational complexity is further reduced by constructing a sparse sequential approximate representation to the posterior process.

2 Processing Scatterometer Data

Scatterometers are commonly used to retrieve wind vectors over ocean surfaces. Current methods of transforming the observed values (scatterometer data, denoted as vector $\boldsymbol{s}$ or $\boldsymbol{s}_i$ at a given spatial location) into wind fields can be split into two phases: local wind vector retrieval and ambiguity removal [Stoffelen and Anderson 1997a] where one of the local solutions is selected as the true wind vector. Ambiguity removal often uses external information, such as a Numerical Weather Prediction (NWP) forecast of the expected wind field at the time of the scatterometer observations. We are seeking a method of wind field retrieval which does not require external data.

In this paper we use a mixture density network (MDN) [Bishop 1995] to model the conditional dependence of the local wind vector $\boldsymbol{z}_i = (u_i, v_i)$ on the local scatterometer observations $\boldsymbol{s}_i$:

$$p_m(\boldsymbol{z}_i|\boldsymbol{s}_i, \boldsymbol{\omega}) = \sum_{j=1}^{4} \beta_{ij} \phi(\boldsymbol{z}_i|\boldsymbol{c}_{ij}, \sigma_{ij}) \tag{1}$$

where $\boldsymbol{\omega}$ is used to denote the parameters of the MDN, ϕ is a Gaussian distribution with parameters functions of $\boldsymbol{\omega}$ and $\boldsymbol{s}_i$. The parameters of the MDN are determined using an independent training set [Evans *et al.* 2000] and are considered known in this application. The MDN which has four Gaussian component densities captures the ambiguity of the inverse problem.

In order to have a global model from the localised wind vectors, we have to combine them. We use a zero-mean vector GP to link the local inverse models [Nabney *et al.* 2000]:

$$q(\underline{\boldsymbol{z}}) \propto \left(\prod_{i}^{N} \frac{p_m(\boldsymbol{z}_i|\boldsymbol{s}_i, \boldsymbol{\omega}) p(\boldsymbol{s}_i)}{p_G(\boldsymbol{z}_i|\boldsymbol{W}_{0i})} \right) p_0(\underline{\boldsymbol{z}}|\boldsymbol{W}_0) \tag{2}$$

where $\underline{\boldsymbol{z}} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^T$ is the concatenation of the local wind field components, $\boldsymbol{W}_0 = \{\boldsymbol{W}_0(x_i, x_j)\}_{ij=1,\ldots,N}$ is the prior covariance matrix for the vector $\underline{\boldsymbol{z}}$ (dependent on the spatial location of the wind vectors), and p_G is p_0 marginalised at z_i, a zero-meaned Gaussian with covariance $\boldsymbol{W}_{0i}$. The choice of the kernel function $\boldsymbol{W}_0(x, y)$ fully specifies our prior beliefs about the model. Notice also that for any given location we have a *two-dimensional* wind vector, thus the output of the kernel function *is a* 2×2 *matrix*, details can be found in [Nabney *et al.* 2000]. The link between two different wind field directions is made through the kernel function – the larger the kernel value, the stronger the "coupling" between the two corresponding wind fields is. The prior Gaussian process is tuned carefully to represent features seen in real wind fields.

Since all quantities involved are Gaussians, we could, *in principle*, compute the resulting probabilities analytically, but this computation is *practically* intractable: the number of mixture elements from $q(\underline{\mathbf{z}})$ is 4^N, extremely high even for moderate values of N. Instead, we will apply the online approximation of [Csató and Opper 2001] to have a jointly Gaussian approximation to the posterior at all data points. However, we know that the posterior distribution of the wind field given the scatterometer observations is multi-modal, with in general two dominating and well separated modes. We might thus expect that the online implementation of the Gaussian process will track one of these posterior modes. Results show that this is indeed the case, although the order of the insertion of the local observations appears to be important.

3 Online learning for the vector Gaussian Process

Gaussian processes belong to the family of Bayesian [Bernardo and Smith 1994] models. However, contrary to the finite-dimensional case, here the "model parameters" are continuous: the GP priors specify a Gaussian distribution over a function space. Due to the vector GP, the kernel function $\boldsymbol{W}_0(x, y)$ is a 2×2 matrix, specifying the pairwise cross-correlation between wind field components at different spatial positions.

Simple moments of GP posteriors (which are usually non Gaussian) have a parametrisation in terms of the training data [Opper and Winther 1999] which resembles the popular kernel-representation [Kimeldorf and Wahba 1971]. For all spatial locations x the mean and covariance function of the vectors $\boldsymbol{z}_x \in \mathbb{R}^2$ are represented as

$$\begin{aligned} \langle \boldsymbol{z}_x \rangle &= \sum\nolimits_{i=1}^{N} \boldsymbol{W}_0(x, x_i) \cdot \boldsymbol{\alpha}_{\boldsymbol{z}}(i) \\ \operatorname{cov}(\boldsymbol{z}_x, \boldsymbol{z}_y) &= \boldsymbol{W}_0(x, y) + \sum\nolimits_{i,j=1}^{N} \boldsymbol{W}_0(x, x_i) \cdot \boldsymbol{C}_{\boldsymbol{z}}(ij) \cdot \boldsymbol{W}_0(x_j, y) \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha}_{\boldsymbol{z}}(1), \boldsymbol{\alpha}_{\boldsymbol{z}}(2), \ldots, \boldsymbol{\alpha}_{\boldsymbol{z}}(N)$ and $\{\boldsymbol{C}_{\boldsymbol{z}}(ij)\}_{i,j=1,N}$ are parameters which will be updated sequentially by our online algorithm. Before doing so, we will (for numerical convenience) represent the vectorial process by a scalar process with twice the number of observations, i.e. we set

$$\langle \boldsymbol{z}_x \rangle = \begin{bmatrix} \langle f_{x^u} \rangle \\ \\ \langle f_{x^v} \rangle \end{bmatrix} \quad \text{and} \quad \boldsymbol{W}_0(x, y) = \begin{bmatrix} K_0(x^u, y^u)\ K_0(x^u, y^v) \\ \\ K_0(x^v, y^u)\ K_0(x^v, y^v) \end{bmatrix} \tag{4}$$

and write (ignoring the superscripts)

$$\begin{aligned} \langle f_x \rangle &= \sum\nolimits_{i=1}^{2N} K_0(x, x_i) \alpha(i) \\ \operatorname{cov}(f_x, f_y) &= K_0(x, y) + \sum\nolimits_{i,j=1}^{2N} K_0(x, x_i) C(ij) K_0(x_j, y) \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{2N}]^T$ and $\boldsymbol{C} = \{C(ij)\}_{i,j=1,\ldots,2N}$ are rearrangements of the parameters from eq. (3).

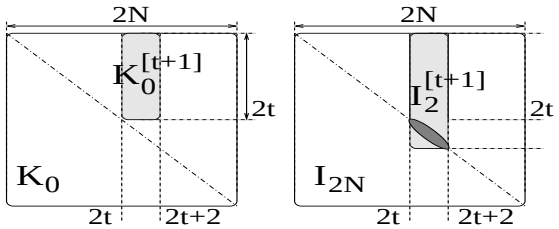


Fig. 1. Illustration of the elements used in the update eq. (6).

The online approximation for GP learning [Csató and Opper 2001] approximates the posterior by a Gaussian at every step. For a new observation $\mathbf{s}_{t+1}$, the previous approximation to the posterior $q_t(\underline{\mathbf{z}})$ together with a local "likelihood" factor (from eq. (2))

$$\frac{p_m(\mathbf{z}_{t+1}|\mathbf{s}_{t+1},\boldsymbol{\omega})p(\mathbf{s}_{t+1})}{p_G(\mathbf{z}_{t+1}|\boldsymbol{W}_{0,t+1})}$$

are combined into a new posterior using Bayes rule. Computing its mean and covariance enable us to create an updated Gaussian approximation $q_{t+1}(\underline{\mathbf{z}})$ at the next step. $\hat{q}(\underline{\mathbf{z}}) = q_{N+1}(\underline{\mathbf{z}})$ is the final result of the online approximation. This process can be formulated in terms of updates for the parameters $\boldsymbol{\alpha}$ and $\mathbf{C}$ which determine the mean and covariance:

$$\begin{aligned}\boldsymbol{\alpha}_{t+1} &= \boldsymbol{\alpha}_t + \boldsymbol{\nu}_{t+1}\frac{\partial \ln g(\langle \mathbf{z}_{t+1}\rangle)}{\partial\langle \mathbf{z}_{t+1}\rangle} \\ \mathbf{C}_{t+1} &= \mathbf{C}_t + \boldsymbol{\nu}_{t+1}\frac{\partial^2 \ln g(\langle \mathbf{z}_{t+1}\rangle)}{\partial\langle \mathbf{z}_{t+1}\rangle^2}\boldsymbol{\nu}_{t+1}^T\end{aligned} \quad \text{with} \quad \boldsymbol{\nu}_{t+1} = \mathbf{C}_t\mathbf{K}_0^{[t+1]} + \mathbf{I}_2^{[t+1]} \tag{6}$$

with elements $\mathbf{K}_0^{[t+1]}$ and $\mathbf{I}_2^{[t+1]}$ are shown in Fig. 1 and

$$g(\langle \mathbf{z}_{t+1}\rangle) = \left\langle \frac{p_m(\mathbf{z}_{t+1}|\mathbf{s}_{t+1},\boldsymbol{\omega})p(\mathbf{s}_{t+1})}{p_G(\mathbf{z}_{t+1}|\boldsymbol{W}_{0,t+1})}\right\rangle_{q_t(\mathbf{z}_{t+1})} \tag{7}$$

and $\langle \mathbf{z}_{t+1}\rangle$ is a vector, implying vector and matrix quantities in (6). Function $g(\langle \mathbf{z}_{t+1}\rangle)$ is easy to compute analytically because it just requires the two dimensional marginal distribution of the process at the observation point $\mathbf{s}_{t+1}$. Fig. 2 shows the results of the online algorithm applied on a sample wind field, details can be found in the figure caption.

3.1 Obtaining sparsity in Wind Fields

Each time-step the number of nonzero parameters will be increased in the update equation. This forces us to use a further approximation which reduces the number of supporting examples in the representations eq. (5) to a smaller set of basis vectors. Following our approach in [Csató and Opper 2001] we remove the

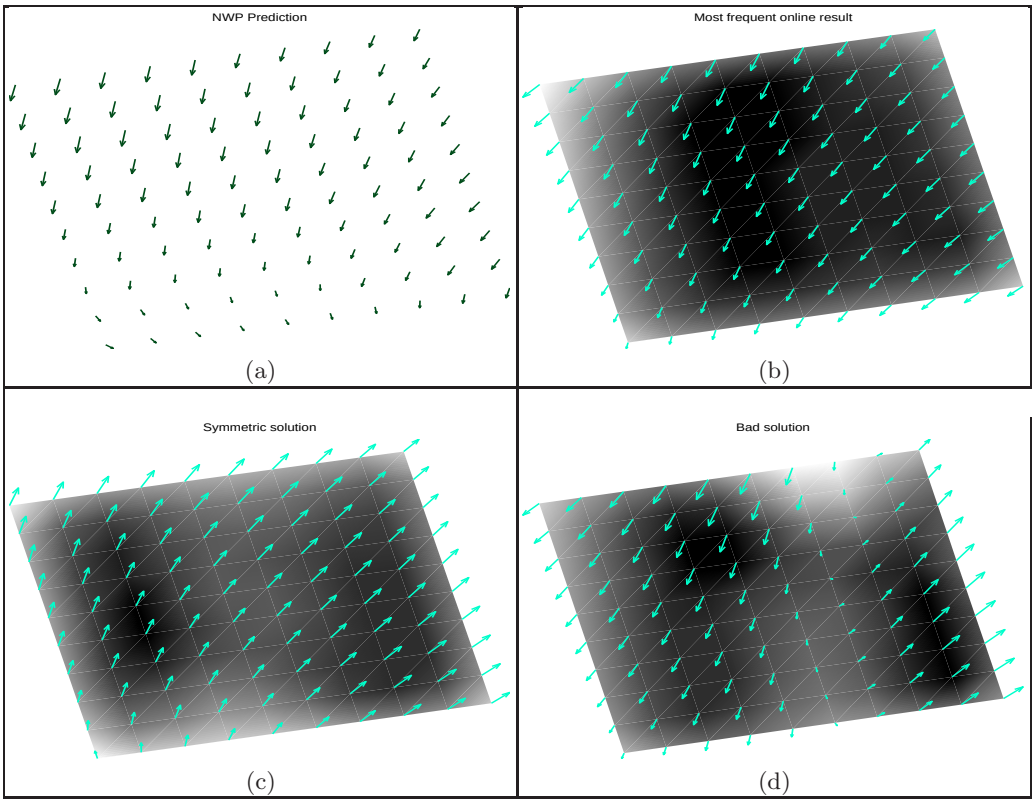


Fig. 2. The NWP wind field estimation (a), the most frequent (b) and the second most frequent (c) online solution together with a bad solution. The assessment of good/bad solution is based on the value of the relative weight from Section 3.2. The gray-scale background indicates the model confidence (Bayesian error-bars) in the prediction, darker shade meaning more confidence.

last data element when a certain score (defined by the feature space geometry associated to the kernel $\mathbf{K}_0$) suggests that the approximation error is small. The remaining parameters are readjusted to partly compensate for the removal as:

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \boldsymbol{\alpha}^{(t)} - \mathbf{Q}^* \mathbf{q}^{*(-1)} \boldsymbol{\alpha}^* \\ \hat{\mathbf{Q}} &= \mathbf{Q}^{(t)} - \mathbf{Q}^* \mathbf{q}^{*(-1)} \mathbf{Q}^{*\mathsf{T}} \\ \hat{\mathbf{C}} &= \mathbf{C}^{(t)} + \mathbf{Q}^* \mathbf{q}^{*(-1)} \mathbf{c}^* \mathbf{q}^{*(-1)} \mathbf{Q}^{*\mathsf{T}} - \mathbf{Q}^* \mathbf{q}^{*(-1)} \mathbf{C}^{*\mathsf{T}} - \mathbf{C}^* \mathbf{q}^{*(-1)} \mathbf{Q}^{*\mathsf{T}}\end{aligned} \tag{8}$$

where $\mathbf{Q}^{-1} = \{K_0(x_i, x_j)\}_{ij=1,\ldots,2N}$ is the inverse of the Gram matrix, the elements being shown in Fig. 3 ($\boldsymbol{\alpha}^*$, $\mathbf{q}^*$ and $\mathbf{C}^*$ are two-by-two matrices).

The presented update is optimal in the sense that the posterior means of the process at data locations are not affected by the approximation [Csató and Opper 2001]. The change of the mean at the location to be deleted is used as a

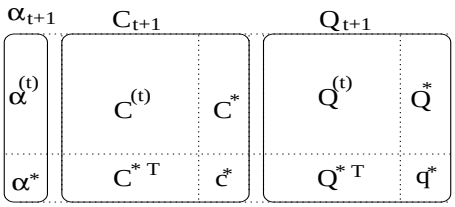
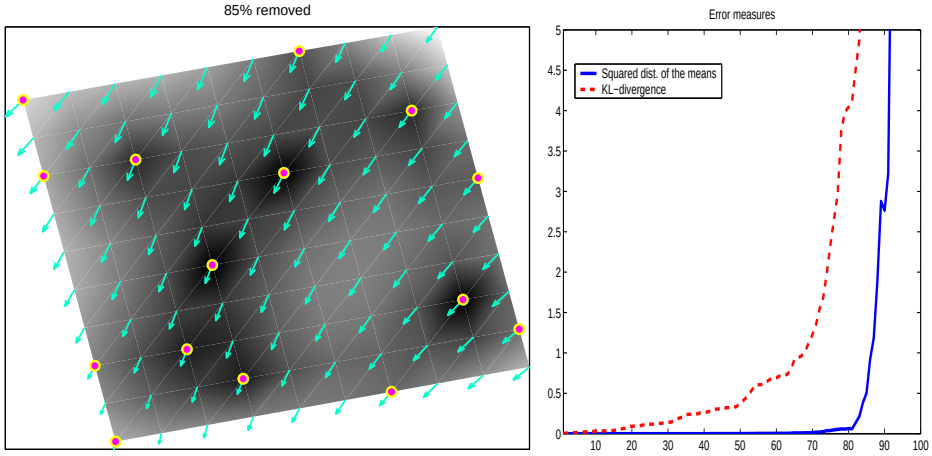


Fig. 3. Decomposition of model parameters for the update equation (8).



(a) (b)

Fig. 4. (a) The predicted wind fields when 85% of the modes has been removed (from Fig. 2). The prediction is based only on basis vectors (circles). The model confidence is higher at these regions. (b) The difference between the full solution and the approximations using the squared difference of means (continuous line) and the KL-distance (dashed line) respectively.

score which measures the loss. This change is (again, very similar to the results from [Csató and Opper 2001]) measured using the score $\varepsilon = \|(\mathbf{q}^*)^{-1}\boldsymbol{\alpha}^*\|$ (the parameters of the vector GP can have any order, we can compute the score for every spatial location).

Removing the data locations with low score sequentially leaves only a small set of so-called *basis points* upon which all further prediction will depend.

Our preliminary results are promising: Fig. 4 shows the resulting wind field if 85 of the spatial knots are removed from the presentation eq. (5). On the right-hand side the evolution of the KL-divergence and the sum-squared errors in the means between the vector GP and a trimmed GP using eq. (8) are shown. as a function of the number of deleted points. Whilst the approximation of the posterior variance decays fast, the the predictive mean is fairly reliable against deleting.

3.2 Measuring the Relative Weight of the Approximation

An exact computation of the posterior would lead to a multi-modal distribution of wind fields at each data-point. This would correspond to a mixture of GPs as a posterior rather than to a single GP that is used in our approximation. If the individual components of the mixture are well separated, we may expect that our online algorithm will track modes with significant underlying probability mass to give a relevant prediction. However, this will depend on the actual sequence of data-points that are visited by the algorithm. To investigate the variation of our wind field prediction with the data sequence, we have generated many random sequences and compared the outcomes based on a simple approximation for the relative mass of the multivariate Gaussian component.

Assuming an online solution of the marginal distribution $(\hat{\underline{z}}, \hat{\mathbf{\Sigma}})$ at a separated mode, we have the posterior at the local maximum expressed:

$$q(\hat{\underline{z}}) \propto \gamma_l \, (2\pi)^{-2N/2} \, |\hat{\mathbf{\Sigma}}|^{-1/2} \tag{9}$$

with $q(\hat{\underline{z}})$ from eq. (2), γ_l *the weight of the component* of the mixture to which our online algorithm had converged, and we assume the local curvature is also well approximated by $\hat{\mathbf{\Sigma}}$.

Having two different online solutions $(\hat{\underline{z}}_1, \hat{\mathbf{\Sigma}}_1)$ and $(\hat{\underline{z}}_1, \hat{\mathbf{\Sigma}}_1)$, we find from eq (9) that the proportion of the two weights is given by

$$\frac{\gamma_1}{\gamma_2} = \frac{q(\hat{\underline{z}}_1)|\hat{\mathbf{\Sigma}}_1|^{1/2}}{q(\hat{\underline{z}}_2)|\hat{\mathbf{\Sigma}}_2|^{1/2}} \tag{10}$$

This helps us to estimate, up to an additive constant, the "relative weight" of the wind field solutions, helping us to assess the quality of the approximation we arrived at. Results, using multiple runs on a wind field data confirm this expectation, the correct solution (Fig. 2.b) has large value and high frequency if doing multiple runs.

4 Discussion

In the wind field example the online and sparse approximation allows us to tackle much larger wind fields than previously possible. This suggests that we will be able to retrieve wind fields using only scatterometer observations, by utilising all available information in the signal.

Proceeding with the removal of the basis points, it would be desirable to have an improved update for the vector GP parameters that leads to a better estimation of the posterior kernel (thus of the Bayesian error-bars).

At present we obtain different solution for different ordering of the data. Future work might seek to build an adaptive classifier that works on the family of online solutions and utilising the relative weights.

However, a more desirable method would be to extend our online approach to mixtures of GPs in order to incorporate the multi-modality of the posterior process in a principled way.

5 Acknowledgement

This work was supported by EPSRC grant no. GR/M81608.

References

[Bernardo and Smith 1994] Bernardo, J. M. and A. F. Smith (1994). *Bayesian Theory*. John Wiley & Sons.

[Bishop 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York, N.Y.: Oxford University Press.

[Cressie 1991] Cressie, N. A. (1991). *Statistics for Spatial Data*. New York: Wiley.

[Csató and Opper 2001] Csató, L. and M. Opper (2001). Sparse representation for Gaussian process models. In T. K. Leen, T. G. Dietrich, and V. Tresp (Eds.), *NIPS*, Volume 13. The MIT Press. http://www.ncrg.aston.ac.uk/Papers.

[Evans *et al.* 2000] Evans, D. J., D. Cornford, and I. T. Nabney (2000). Structured neural network modelling of multi-valued functions for wind retrieval from scatterometer measurements. *Neurocomputing Letters 30*, 23–30.

[Kimeldorf and Wahba 1971] Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic. 33*, 82–95.

[Nabney *et al.* 2000] Nabney, I. T., D. Cornford, and C. K. I. Williams (2000). Bayesian inference for wind field retrieval. *Neurocomputing Letters 30*, 3–11.

[Offiler 1994] Offiler, D. (1994). The calibration of ERS-1 satellite scatterometer winds. *Journal of Atmospheric and Oceanic Technology 11*, 1002–1017.

[Opper 1998] Opper, M. (1998). A Bayesian approach to online learning. See Saad [1998], pp. 363–378.

[Opper and Winther [1999] Opper, M. and O. Winther (1999). Gaussian processes and SVM: Mean field results and leave-one-out estimator. In A. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans (Eds.), *Advances in Large Margin Classifiers*, pp. 43–65. Cambridge, MA: The MIT Press.

[Saad [1998] Saad, D. (1998). *On-Line Learning in Neural Networks*. Cambridge Univ. Press.

[Stoffelen and Anderson [1997a] Stoffelen, A. and D. Anderson (1997a). Ambiguity removal and assimiliation of scatterometer data. *Quarterly Journal of the Royal Meteorological Society 123*, 491–518.

[Stoffelen and Anderson [1997b] Stoffelen, A. and D. Anderson (1997b). Scatterometer data interpretation: Estimation and validation of the transfer function CMOD4. *Journal of Geophysical Research 102*, 5767–5780.