Bayesian analysis of the scatterometer wind retrieval inverse problem: some new approaches

Dan Cornford, Lehel Csató, David J. Evans, Manfred Opper

Neural Computing Research Group, Aston University, Birmingham B4 7ET, UK

Summary. The retrieval of wind vectors from satellite observed radar backscatter can be seen as a non-linear inverse problem. A common approach to solving inverse problems is the Bayesian framework: to infer the posterior distribution of the latent variables of interest given the observations, a model relating the observations to the latent variables, and a prior distribution over the latent variables. In this paper we show how Gaussian process priors can be used in a variety of retrieval methods, using local forward (observation) models and direct inverse models. We present an enhanced Markov Chain Monte Carlo method to sample from the resulting multi-modal posterior distribution.
We go on to show how the computational complexity of the inference can be controlled using sparse, sequential Bayesian learning for Gaussian processes. This helps to overcome the most serious barrier to the use of fully probabilistic, Gaussian processes methods in remote sensing inverse problems, where the size of the data set can become prohibitively large. We contrast the sampling results with the approximations found using the sparse sequential Gaussian process algorithm.

1. Introduction

Satellite borne scatterometers have been in existence for several decades; retrieving wind vectors from scatterometer observations is a complex inverse problem. In brief, scatterometers emit a pulse of micro-wave radiation which travels largely unmodified through the atmosphere. When it meets the ocean surface it interacts with water waves of a similar wavelength ($\sim 5\ cm$ for C-band radars): the strength of the backscattered signal depending on the magnitude and relative orientation of the water waves to the incoming radiation. The water wave field at these short (centimetre) wavelengths is driven by the instantaneous surface wind stress, which in turn is related to the near surface (usually taken at 10 m height) wind vector. The physics of the problem is complex and not fully understood (Janssen et al., 1998). So far, most solutions to this inverse problem have been based on inverting an empirical sensor model for the scatterometer observation process (Stoffelen and Anderson, 1997a), often using look up tables (Offiler, 1994). In this paper we propose a principled, statistically based retrieval method for the scatterometer inverse problem, and demonstrate how an innovative Bayesian sequential learning algorithm can be used to significantly speed up the solution of this inverse problem. We apply our methods to data assimilation, where we have an informative prior estimate of the state (taken e.g. from a dynamical model) and data retrieval with a zero mean prior.

The paper is organised as follows: in Section 2 the methods for obtaining a local solution to the inverse problem are shown. Section 3 shows how the prior vector Gaussian process models are constructed and Section 4 how the priors are combined with the local solutions to perform a field-wise solution to the inverse problem. The results are shown and discussed in Section 5 while conclusions are drawn in Section 6.

1.1. The aim of scatterometry

Obtaining wind vectors over the ocean is important to numerical weather prediction (NWP) since the ability to produce a forecast of the future state of the atmosphere depends critically on knowing the current state accurately (Haltiner and Williams, 1980). However, the observation network over the oceans (particularly in the Southern Hemisphere) is very limited (Daley, 1991). Thus it is hoped that the global coverage of ocean wind vectors provided by satellite-borne scatterometers (Offiler, 1994) will improve the accuracy of weather forecasts by providing better initial conditions for NWP models (Lorenc et al., 1993). The scatterometer data also offer the potential of improved wind climatologies over the oceans (Levy, 1994) and the possibility of studying, at high resolution, interesting meteorological features such as cyclones (Dickinson and Brown, 1996).

1.1.1. Scatterometers

This study uses scatterometer data from the ERS-2 satellite; the on-board vertically polarised microwave radar operates at 5.3 GHz and measures the backscatter from gravity-capillary waves on the ocean surface of $\sim 5\ cm$ wavelength. Backscatter from the ocean surface is measured by the normalised radar cross section, generally denoted by s^o, and has units of decibels. A 500-km wide swathe is swept by the satellite to the right of the track of its polar orbit. There are 19 cells sampled across the swathe, and each cell has dimensions of roughly 50 by 50 km, which implies that there is some overlap between cells.

Each cell is sampled from three different directions by the 'fore', 'mid', and 'aft' beams, giving a triplet, $\boldsymbol{s^o} = (s_f^o, s_m^o, s_a^o)$. This triplet $\boldsymbol{s^o}$, together with the incidence and azimuth angles of the beams (which vary across the swathe), is related to the average wind vector $\boldsymbol{v}$ within the cell (Offiler, 1994). We assume that any unmodelled effects are largely related to wind speed and thus their impact is implicitly included in the empirical models which have been developed (Cornford et al., 2001; Bullen et al., 2003). Other geophysical parameters such as rain and sea ice are also

believed to have a small effect on the backscatter (Stoffelen, 1998); however, since we have no independent measurements of these phenomena, they are treated as additional noise sources in this paper.

1.2. Bayesian framework

The inverse problem to be solved is the retrieval of the wind field, $\boldsymbol{V} = \{\boldsymbol{v}_i\}_{i=1,N}$, given the scatterometer observations, $\boldsymbol{S^o} = \{\boldsymbol{s^o_i}\}_{i=1,N}$, where N is the number of observations. Since the processes involved include noise, it is necessary to retrieve the conditional probability density function, $p(\boldsymbol{V} \mid \boldsymbol{S^o})$. In the usual way for Bayesian approaches to the solution of inverse problems (Tarantola, 1987) this posterior probability is written:

$$p(\boldsymbol{V} \mid \boldsymbol{S^o}) = \frac{p(\boldsymbol{S^o} \mid \boldsymbol{V})p(\boldsymbol{V})}{p(\boldsymbol{S^o})} \propto \left(\prod_i p(\boldsymbol{s^o_i} \mid \boldsymbol{v}_i) \right) p(\boldsymbol{V}) \; . \tag{1}$$

The likelihood $p(\boldsymbol{S^o} \mid \boldsymbol{V})$ is assumed to factorise as $\prod_i p(\boldsymbol{s^o_i} \mid \boldsymbol{v}_i)$, that is the errors on the scatterometer observations are assumed independent. For the estimation of the posterior distribution the normalising constant in the denominator needs to computed, which is seldom possible analytically. We show two approaches to the inference of $p(\boldsymbol{V} \mid \boldsymbol{S^o})$ using forward models, based on hybrid Monte Carlo sampling and sequential Bayesian learning.

Using the scaled likelihood method (Morgan and Boulard, 1995), repeating the application of Bayes theorem to the local likelihood, it is possible to rewrite eq. (1) to use direct local inverse models, $p(\boldsymbol{v}_i \mid \boldsymbol{s^o_i})$, by noting that:

$$p(\boldsymbol{s^o_i} \mid \boldsymbol{v}_i) \propto \frac{p(\boldsymbol{v}_i \mid \boldsymbol{s^o_i})}{p(\boldsymbol{v}_i)}. \tag{2}$$

where we observed the $\boldsymbol{S^o}$ values thus we can discard the proportionality constant. Replacing the product term in eq. (1) using eq. (2) gives:

$$p(\boldsymbol{V} \mid \boldsymbol{S^o}) \propto \left(\prod_i \frac{p(\boldsymbol{v}_i \mid \boldsymbol{s^o_i})}{p(\boldsymbol{v}_i)} \right) p(\boldsymbol{V}) \; . \tag{3}$$

We show a solution to eq. (3) using a multi-modal sampling method.

In the discussion to this point $p(\boldsymbol{V})$ has remained undefined. A natural probabilistic representation for wind fields, which links the local models, is to use a vector Gaussian process:

$$p(\boldsymbol{V}) = \frac{1}{(2\pi)^{\frac{N}{2}} \, |K_{\boldsymbol{v}}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{V}' K_{\boldsymbol{v}}^{-1} \boldsymbol{V} \right) \; , \tag{4}$$

where $K_{\boldsymbol{v}}$ is the joint covariance matrix for $\boldsymbol{V}$. The parameterisation of the prior wind field model is discussed in Section 3.

2. Local solution

The initial problem which must be addressed is the local (that is cell-wise) inversion of the observation operator. As discussed in section 1 the sensor or forward model, which maps the state variables that are to be recovered ($\boldsymbol{v}$) into the observables ($\boldsymbol{s^o}$) is not perfectly known. Physics based models such as VIERS1 (Janssen et al., 1998) have been shown to provide reasonable approximations to observed data, and thus empirical approaches have been developed which exploit some of the prior knowledge from the physics of the observation processes by imposing a truncated Fourier model structure (Stoffelen and Anderson, 1997b). In Section 2.1 we show how neural networks can be used to improve these forward models by providing more flexible functional forms, while Section 2.2 describes the use of mixture density networks to construct local direct inverse models.

A significant issue when dealing with scatterometer data comes from recognising that this data provides information on the wind vector over a 50 by 50 km cell over the ocean. There are no other measurements which can be directly made on $\boldsymbol{v}$ at this scale. At best there may be a single buoy measurement somewhere within the region, however there is no guarantee this will be representative. The best approximation to the average wind vector can be derived from Numerical Weather Prediction (NWP) models. These pseudo-observations are numerical forecasts of the dynamics of the atmosphere which are very complex models, integrating coupled partial differential equations with a significant number of parameterisations, and incomplete initial and boundary conditions (Haltiner and Williams, 1980). Thus, forecasts from NWP models will themselves be in error, and they must be borne in mind when using NWP data: this should not be regarded as truth.

The training of all the local models described in this paper was conducted on a set of collocated scatterometer and NWP wind vectors, which are extracted from the European Centre for Medium Range Weather Forecasting NWP model. All results are reported on a similarly obtained test set using the Met Office NWP model for completely different time periods making the training and test sets independent.

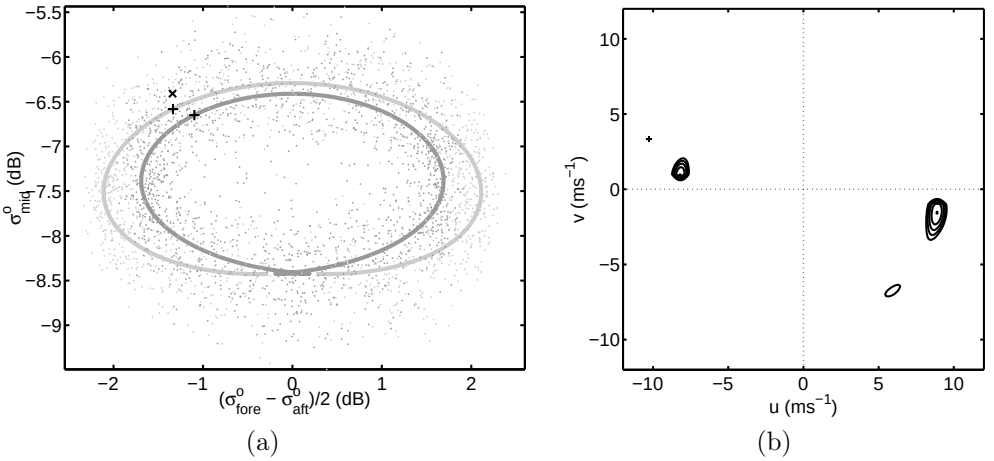

Fig. 1. (a) An approximate cross section through the 3D cone which de£nes the forward model. The lighter part is the upwind segment, the darker the downwind segment. A satellite observation is plotted as x, and corresponding closest points on the model manifold, are marked $+$. Also shown are samples from the noise distribution along the manifold (small grey dots), this giving an idea of instrument noise. (b) A contour plot of the local conditional probability density function, $\log[p(\boldsymbol{s^o} \mid \boldsymbol{v})]$, as a function of $\boldsymbol{v}$. The NWP wind vector is marked by the cross and the two humps correspond to the two projections in sub-£gure (a).

2.1. Forward models

In recent work (Bullen et al., 2003) we developed a forward model based on a combination of a radial basis function network and a truncated Fourier series:

$$\boldsymbol{s^o} = a_0 + a_1 \cos(\chi) + a_2 \cos(2\chi) + a_3 \cos(3\chi) + a_4 \cos(4\chi) \ , \tag{5}$$

where $a_0, a_1 \ldots a_4$ are the outputs from the radial basis function network with inputs wind speed and beam incidence angle, and χ is the wind direction relative to the satellite azimuth angle. This was trained in a Bayesian framework which accounted for "input" noise on $\boldsymbol{v}$ and utilised a radial basis function network to estimate the noise variance, again trained in a Bayesian framework (Bishop and Qazaz, 1997). The forward model provides a local estimate of $p(\boldsymbol{s^o} \mid \boldsymbol{v})$, and can locally be inverted (using non-linear optimisation) to retrieve $\boldsymbol{v}$, however care must be taken due to the presence of multiple local solutions, resulting from the Fourier form of the forward model and the observation noise, as illustrated in Fig. 1.

This ambiguity can be understood when considering looking at waves side on (these small gravity waves are almost totally symmetric). With a static image it is almost impossible to tell the direction of motion; this intuition carries over well into the microwave region of the spectrum. When observation noise is also taken into account there are typically between 2 and 4 local solutions, thus any inversion process must take this into account: simple gradient based optimisation will only discover a single mode. In practice multiple initialisations are used to determine all $\boldsymbol{v}$ which maximise $p(\boldsymbol{s^o} \mid \boldsymbol{v})$.

For practical applications where many thousands of scatterometer observations must be processed quickly, the non-linear optimisation from multiple starts is prohibitively expensive, and only provides the most probable $\boldsymbol{v}$, not $p(\boldsymbol{v} \mid \boldsymbol{s^o})$, which is desired. In some situations it is possible to design direct local inverse models (Ward and Redfern, 1999) but in this case the presence of multiple solutions for a given $\boldsymbol{s^o}$ observation precludes a simple approach.

2.2. Direct inverse models

To model the multi-modal $p(\boldsymbol{v} \mid \boldsymbol{s^o})$ we have employed mixture density networks (Bishop, 1995). In previous work (Evans, 2001) these were trained directly on NWP data, however it is almost impossible in this case to separate the contribution of the uncertainty deriving from the use of NWP estimates and the uncertainty coming from the observation noise. In this paper we use a mixture density network trained on data generated from the forward model (Section 2.1) with simulated observational noise.

This direct inverse model efficiently provides an estimate of $p(\boldsymbol{v} \mid \boldsymbol{s^o})$ using a single forward propagation, which is a significant speed improvement over the nonlinear optimisation based methods and look up table based methods (as currently used in the UK Met Office) which only return a single value. Using mixture models will approximate the *distribution* of $\boldsymbol{v}$.

3. Vector Gaussian process priors

A stationary vector Gaussian process model is used to represent wind fields, based on the decomposition of a vector field into purely divergent and purely rotational flow, known as Helmholtz's theorem (Daley, 1991). We do not give details here, but the full development can be found in (Cornford, 1998). Essentially, a scalar modified Bessel covariance function † based Gaussian process (Handcock and Wallis, 1994) was applied independently to both the stream function

†the modified Bessel covariance function is also referred to as a Matérn covariance function.

and velocity potential, rather than directly to the wind vector components. The wind vector components can be written in terms of derivatives of the stream function and velocity potential, so the covariance functions for the wind vector components are computed in terms of second order derivatives of the stream function and velocity potential covariances (Cornford, 1998). This allows control over the ratio of divergence to vorticity in the resulting wind field and automatically produces valid, positive definite, joint covariance matrices $K_{\boldsymbol{v}}$ for the wind vector components.

The parameters of the zero mean modified Bessel covariance function are learnt using a sector of the North Atlantic from $52.5°W, 40°N$ to $10°W, 60°N$, since this is the most reliably observed ocean region. A large amount of data from the European Centre for Medium range Weather Forecasting (ECMWF) is used; this consists of gridded analysis data on a regular 2.5 degree latitude-longitude grid. In order to obtain reliable climatological estimates of the parameters in the wind field model, three complete years of data (1995–97) were used. To account for seasonality, a separate prior wind field model is developed for each month, using the three years data, thus the prior is strictly tuned to the North Atlantic region. The prior distribution of the kernel parameters are chosen by expert assessment. The maximum *a posteriori* probability values of the parameters are used in subsequent experiments. These values suggested that the wind fields are once differentiable. Thus the modified Bessel covariance function simplifies to a polynomial-exponential covariance function which has the form:

$$C(r) = E^2 \left(1 + \frac{r}{L} + \frac{r^2}{3L^2}\right) \exp\left(-\frac{r}{L}\right) + \eta^2 \tag{6}$$

where r is the separation distance of two points, L is a characteristic length scale parameter, E^2 is the process variance and η^2 is the noise variance. This form is much quicker to compute and is used for this reason. In the text we treat the maximum *a posteriori* probability parameters of the Gaussian process as known and fixed, and do not explicitly condition on them. The covariance matrix, $K_{\boldsymbol{v}}$, of eq. (4) is obtained using Helmholtz's theorem and the covariance function (6) for the stream function and velocity potential independently (Cornford, 1998).

4. Bayesian retrieval

Section 1.2 describes the theoretical environment in which this work is cast. In this section we introduce the practical methods which are applied to approximate the full posterior distribution from eq. (3) or eq. (1), based on sampling and sequential Bayesian learning, in contrast to traditional maximum *a posteriori* probability approaches used in data assimilation.

4.1. Sampling

In this section we turn our attention to drawing samples from the posterior distribution of the direct inverse model, eq. (3). The likelihood term involves a mixture model, thus the posterior cannot be computed analytically due to the factorial growth in complexity. We apply Markov Chain Monte Carlo (MCMC) techniques in order to explore the posterior distribution. MCMC techniques are powerful when applied to such distributions as they make no assumptions about the form of the distribution.

Earlier work has shown that the posterior distribution is multi-modal, and typically it is dominated by two well separated modes (Evans, 2001). Sampling from the posterior distribution is thus implemented by an extension of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) incorporating knowledge of the dominating modes located in a mode finding exercise.

Due to the number of local modes an exhaustive search is impractical. Instead, a mixture of probabilistic and deterministic methods are applied. Modes in the posterior distribution are identified by several hundred optimiser searches, where the optimiser is initialised to a random sample from the prior distribution, and run until it converges. The cost function of the optimiser is the negative log posterior from eq. (3), which is the energy function. The solutions are clustered into modes using the Euclidean distance in $\boldsymbol{V}$ and the minimum energy (Evans, 2001). This method identifies several modes. However, as mentioned two of them dominate the posterior distribution, and are approximately symmetric. The remaining modes may be ignored as it has been shown that these are local minima within the basin of attraction of one of the dominating modes (Evans, 2001).

In the Metropolis-Hastings algorithm, a new state $\boldsymbol{V}_{t+1}$ is generated from the previous state, $\boldsymbol{V}_t$, by generating a candidate state $\boldsymbol{V}^*$ from a *proposal distribution*, $q(\boldsymbol{V}^*|\boldsymbol{V}_t)$, and then deciding whether or not to accept the candidate state based on its probability density relative to that of the old state, with respect to the desired invariant distribution eq. (3).

Assume two posterior modes have been identified, labelled $\bar{\boldsymbol{V}}_{\mathbf{1}}$ and $\bar{\boldsymbol{V}}_{\mathbf{2}}$. The vector $\mathbf{J}_{12} = \bar{\boldsymbol{V}}_{\mathbf{2}} - \bar{\boldsymbol{V}}_{\mathbf{1}}$ defines a jump in state space from $\bar{\boldsymbol{V}}_{\mathbf{1}}$ to $\bar{\boldsymbol{V}}_{\mathbf{2}}$. To propose a jump from $\bar{\boldsymbol{V}}_{\mathbf{1}}$ to $\bar{\boldsymbol{V}}_{\mathbf{2}}$, we modify the proposal distribution to incorporate the jump: $q(\boldsymbol{V}^*|\boldsymbol{V}_t + \mathbf{J}_{12})$. To sample within the current mode we set the jump to be zero. Assume we are in mode 1, then a new sample is generated by the following algorithm:

(a) Propose a mode jump with probability 0.5; $\mathbf{J} = 0$ or $\mathbf{J} = \mathbf{J}_{12}$
(b) Propose a new state, $\boldsymbol{V}^*$, by drawing a sample conditionally on the current state $\boldsymbol{V}_t$:

$$\boldsymbol{V}^* = \boldsymbol{V}_t + \mathbf{J} + N(0, \sigma^2 K_{\boldsymbol{v}}). \tag{7}$$

(c) Compute the acceptance probability:

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{V}^*|\boldsymbol{S^o})q((\boldsymbol{V}_t + \mathbf{J})|\boldsymbol{V}^*)}{p(\boldsymbol{V}_t|\boldsymbol{S^o})q(\boldsymbol{V}^*|(\boldsymbol{V}_t + \mathbf{J}))}\right\}. \tag{8}$$

(d) Draw sample α_{test} from uniform distribution $[0, 1]$.
(e) If $\alpha > \alpha_{test}$ then set $\boldsymbol{V}_{t+1} = \boldsymbol{V}^*$; else $\boldsymbol{V}_{t+1} = \boldsymbol{V}_t$.

To ensure that it is possible to jump back from $\bar{\boldsymbol{V}}_{\mathbf{2}}$ we must also propose a jump $\bar{\boldsymbol{V}}_{\mathbf{1}} = \bar{\boldsymbol{V}}_{\mathbf{2}} - \mathbf{J}$ in our sampler scheme with equal probability to the jump from $\bar{\boldsymbol{V}}_{\mathbf{1}}$ to $\bar{\boldsymbol{V}}_{\mathbf{2}}$; this maintains detailed balance for the sampler. The above method is equivalent to combining the proposal distributions by the *mixing* method described in (Brooks, 1999).

The practical advantage of this method is that the mode finding is completed before the sampling commences. This means that specialist mode finding methods may be employed for each particular problem, and that computational effort required to find the modes is not repeated many times during a sampling run. The presented algorithm is similar to existing methods for sampling from multi-modal distributions, such as simulated tempering (Marinari and Parisi, 1992), tempered transitions (Neal, 1995) and mode jumping kernels for MCMC (Tjelmeland and Hegstad, 2001). These all rely on simultaneously identifying modes and sampling. In practice, with the well separated modes in this problem, the existing methods do not converge in any reasonable time, and the initial mode finding is the only computationally feasible method.

If the aim is to efficiently explore each mode individually (or where the problem has been rendered uni-modal by the use of a non-zero mean prior) the hybrid Monte Carlo algorithm (Duane et al., 1987) is used to explore of the uncertainty within the modes.

4.2. Sparse, sequential Bayesian learning

We next turn to our second approach for tackling the intractable computations with the posterior distribution,

$$p(\boldsymbol{V}|\boldsymbol{S^o}) \propto \prod_{i=1}^{N} L_i(\boldsymbol{v}_i)\; p(\boldsymbol{V})\;. \tag{9}$$

Here $L_i(\boldsymbol{v}_i) = P(\boldsymbol{s^o}_i|\boldsymbol{v}_i)$ is shorthand for the likelihood of the i'th observation from either eq. (1) or (3), and we will drop the explicit dependence of the posterior on the observations from now on. Our method is based on an approximation to eq. (9) which allows us to reduce the problem of computing $2N$-dimensional integrals over the wind fields $\boldsymbol{V}$ by a sequence of *two* dimensional integrations.

We use the fact that we can build up the posterior eq. (9) sequentially, by processing the observations at different spatial locations one after the other in an arbitrary sequence. Assuming that we have already processed t data points, the next observation with its likelihood $L_{t+1}(\boldsymbol{v}_{t+1})$ leads to the exact update of the posterior distribution

$$p_{t+1}(\boldsymbol{V}) = \frac{1}{Z_{t+1}}\; L_{t+1}(\boldsymbol{v}_{t+1})\; p_t(\boldsymbol{V})\;, \tag{10}$$

where Z_{t+1} is a normalisation constant. Having observed all N data points, we get $p_N(\boldsymbol{V})$ as the desired full posterior eq. (9).

Initialising the sequential update with the tractable Gaussian prior distribution $p_0(\boldsymbol{V}) = p(\boldsymbol{V})$ from eq. (4), the first observation leads to a non-Gaussian joint density $p_1(\boldsymbol{V})$ for the set of wind vectors $\boldsymbol{V}$. The basic idea of our approximation is to *project* p_1 to a sensibly defined "closest" Gaussian distribution q_1. By using q_1 in place of p_1 for the update eq. (10) we get again a non-Gaussian approximation $\tilde{p}_2(\boldsymbol{V})$ to the true posterior $p_2(\boldsymbol{V})$. We can repeat the projection method to obtain a tractable Gaussian approximation q_2. Performing the projection and update steps for any observation in the sequence, ie.

$$\begin{aligned} \tilde{p}_{t+1}(\boldsymbol{V}) &= \frac{1}{\tilde{Z}_{t+1}}\; L_{t+1}(\boldsymbol{v}_{t+1})\; q_t(\boldsymbol{V}) \\ \tilde{p}_{t+1}(\boldsymbol{V}) &\to q_{t+1}(\boldsymbol{V}) \end{aligned} \tag{11}$$

for $t = 0, \ldots, N-1$ leads to a sequence q_t of Gaussian approximations to the true posteriors p_t. To minimize the instantaneous loss of information in the projection step, we choose q_t to be the closest distribution to $\tilde{p}_t$ in the family $\mathcal{G}$ of Gaussian distributions where closeness is measured by the (nonsymmetric) relative entropy or Kullback-Leibler (KL) divergence (Cover and Thomas, 1991). This means, we set

$$q_t(\boldsymbol{V}) = \underset{q \in \mathcal{G}}{\operatorname{argmin}} \int d\boldsymbol{V}\; \tilde{p}_t(\boldsymbol{V}) \ln \frac{\tilde{p}_t(\boldsymbol{V})}{q(\boldsymbol{V})}\;. \tag{12}$$

Note, that our choice of the order of original and approximating distributions is different from the one used in the *variational approximation* (see e.g. Zoubin and Beal in (Opper and Saad, 2001)) frequently used in the machine learning community. We believe that our choice is more sensible than the reversed one, because the log-ratio in eq. (12) is weighted by the less approximate density. It can be easily shown that the minimization in eq. (12) is equivalent to

choosing the Gaussian density q_t which has the *same first and second moments* as $\tilde{p}$. The idea of approximating a posterior by sequentially *propagating moments* has been discovered independently in different scientific communities and applied to inference problems with Gaussian processes in Csató and Opper (2002).

The only remaining nontrivial task is the computation of the moments, ie. the mean vector and of the covariance matrix for the *non-Gaussian distribution* $\tilde{p}_{t+1}$. However, since all variables $\boldsymbol{v}_i$ except for $\boldsymbol{v}_{t+1}$ appear only in the Gaussian density q_t in eq. (11), the corresponding multidimensional integrals are easily reduced to two-dimensional non-Gaussian ones over the vector $\boldsymbol{v}_{t+1}$. In many cases of interest, such as the Gaussian mixture model specified by the direct inverse model introduced in section 2.2 (Csató et al., 2001), these can be performed analytically.

It should be noted that although we process a single observation at a time, at each instant our method updates an approximate joint Gaussian density of the wind field at *all* observation points. At step t this density has the form

$$q_t(\boldsymbol{V}) \propto \mathcal{L}_t \; p_0(\boldsymbol{V}) \; , \tag{13}$$

where the logarithm of the effective total "Gaussian" likelihood $\mathcal{L}_t$ is quadratic in those windfields $\boldsymbol{V}_t = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_t)$ which have already been processed by the sequential algorithm. A simple "representer" lemma, proved in (Csató and Opper, 2002), shows how the means and the covariances for the set of fields at *all* points can be represented using a $2t$ dimensional vector and a $2t \times 2t$ dimensional matrix which both depend only on the t spatial locations which have been "visited" by the algorithm sofar. This vector and the matrix are the parameters which are processed by the sequential algorithm. It recomputes their elements and enlarges their dimensionalities any time a new data point is added.

The propagation of a Gaussian approximation to the posterior distribution is in the spirit of the *Kalman filter* algorithm (Kalman and Bucy, 1961). There is however a crucial difference which comes from the fact that we do not introduce a *local* approximation to the original likelihood (eg. linearization). The integration in eq. (11) is a *global* smoothing operation which allows us to deal with likelihoods that are nonsmooth or even non-continuous functions of the latent variables.

Although we have dealt with the problem of high dimensional intractable integrals, the major problem for a practical application of this sequential Bayesian procedure to realistic problems is the drastic increase in computational complexity with the number of observations, N. The increase in the size of the matrix used by the algorithm makes an application to large systems infeasible. Hence, a further *sparse* approximation is necessary. The simplest idea would be to *discard a new datapoint* if its influence on the entire posterior process is sufficiently small. This influence could be measured by the change of the approximated posterior density when the datapoint is being processed. To quantify the change, one could use an information theoretic measure like the relative entropy defined in eq. (12). Discarding a substantial fraction of observations would then obviously lead to a slower increase of the matrix size but may on the other hand lead to a substantial loss of information.

With slightly more effort, we can still keep some amount of the information contained in the discarded observations without increasing the size of the matrix. This is achieved by replacing the data likelihood $\mathcal{L}_t$ in eq. (13) by a further *optimized* sparse approximation $\mathcal{L}_t^s$ which is still the exponential of a quadratic form but depends on a small set of observations only. The sparse approximation is performed sequentially together with the projection and update steps of the Bayes on-line algorithm introduced above. Suppose that at step t, we have a sparse approximation to the likelihood $\mathcal{L}_t^s(\boldsymbol{V}^s)$ which depends only on a subset $\boldsymbol{V}^s$ of wind vectors which we call *basis vectors*. The update and projection step applied to a new datapoint leads to an updated Gaussian posterior q_{t+1} and a corresponding effective Gaussian likelihood $\mathcal{L}_t(\boldsymbol{V}^s, \boldsymbol{v}_{t+1})$.

We aim at constructing an approximation to the posterior which is of the form

$$q_{t+1}^s \propto \mathcal{L}_{t+1}^s(\boldsymbol{V}^s) \; P_0(\boldsymbol{V})$$

for which the new likelihood $\mathcal{L}_{t+1}^s$ is again sparse, ie. a function of the basis vectors only. It has to be optimized in order to make q_{t+1} and q_{t+1}^s as close as possible. The optimization is most easily performed when we measure the closeness of densities by the KL divergence

$$\int d\boldsymbol{V} \; q_{t+1}^s(\boldsymbol{V}) \ln \frac{q_{t+1}^s(\boldsymbol{V})}{q_{t+1}(\boldsymbol{V})} \; . \tag{14}$$

Note, that compared to eq. (12) the order of original and approximated distributions have been interchanged. One finds that the optimal likelihood is obtained by replacing $\boldsymbol{v}_{t+1}$ by its *minimal mean square prediction* $\hat{\boldsymbol{v}}_{t+1}(\boldsymbol{V}^s)$ given the basis vectors under the prior distribution. This means that we replace $\boldsymbol{v}_{t+1}$ by its projection on the space spanned by the basis vectors Vb^s, where the metric is induced by the inner product defined by the kernel K. Hence, we get

$$\mathcal{L}_{t+1}^s(\boldsymbol{V}^s) = \mathcal{L}_{t+1}(\boldsymbol{V}, \hat{\boldsymbol{v}}_{t+1}(\boldsymbol{V}^s)) \; . \tag{15}$$

When the minimal KL divergence is smaller than some given tolerance ϵ, we replace q_{t+1} by q_{t+1}^s and the size of the basis set is not increased. Otherwise, we augment the basis vector set by the field $\boldsymbol{v}_{t+1}$. If the number of basis vectors increases beyond a maximally tollerable number, we can apply the sparsification method to discard old basis vectors which achieve a small information score using eq. (15).

There is an inherent problem in the sequential method defined sofar which stems from the dependence of the final prediction on the order in which the observations are processed which is arbitrary. It is thus highly desirable to be able to sweep several times through the data to refine the approximation until convergence is achieved. A naive repetition of the projection and update steps eq. (11) in a second sweep through the data would be highly inconsistent. It would

violate the assumption (which is inherent in the sequential construction of the posterior) that the local likelihood for the field $\boldsymbol{v}_{t+1}$ was not yet included in the posterior.

A method for solving this problem within the general framework of sequential Bayesian methods has been suggested by Minka (2000). In the *expectation propagation* approach, he shows how to approximately delete the information contained in an observation from the posterior before the update / projection step is performed again on this datapoint. For Gaussian process models it has been shown that when this procedure converges, the final approximation is equivalent to the so called TAP (named after Thouless, Anderson and Palmer) approximation (Opper and Winther, 2001). The latter method was originally developed in the field of statistical physics (Mezard et al., 1987) and has been recently applied successfully to approximate inference with a variety of probabilistic data models (for a review see (Opper and Saad, 2001)).

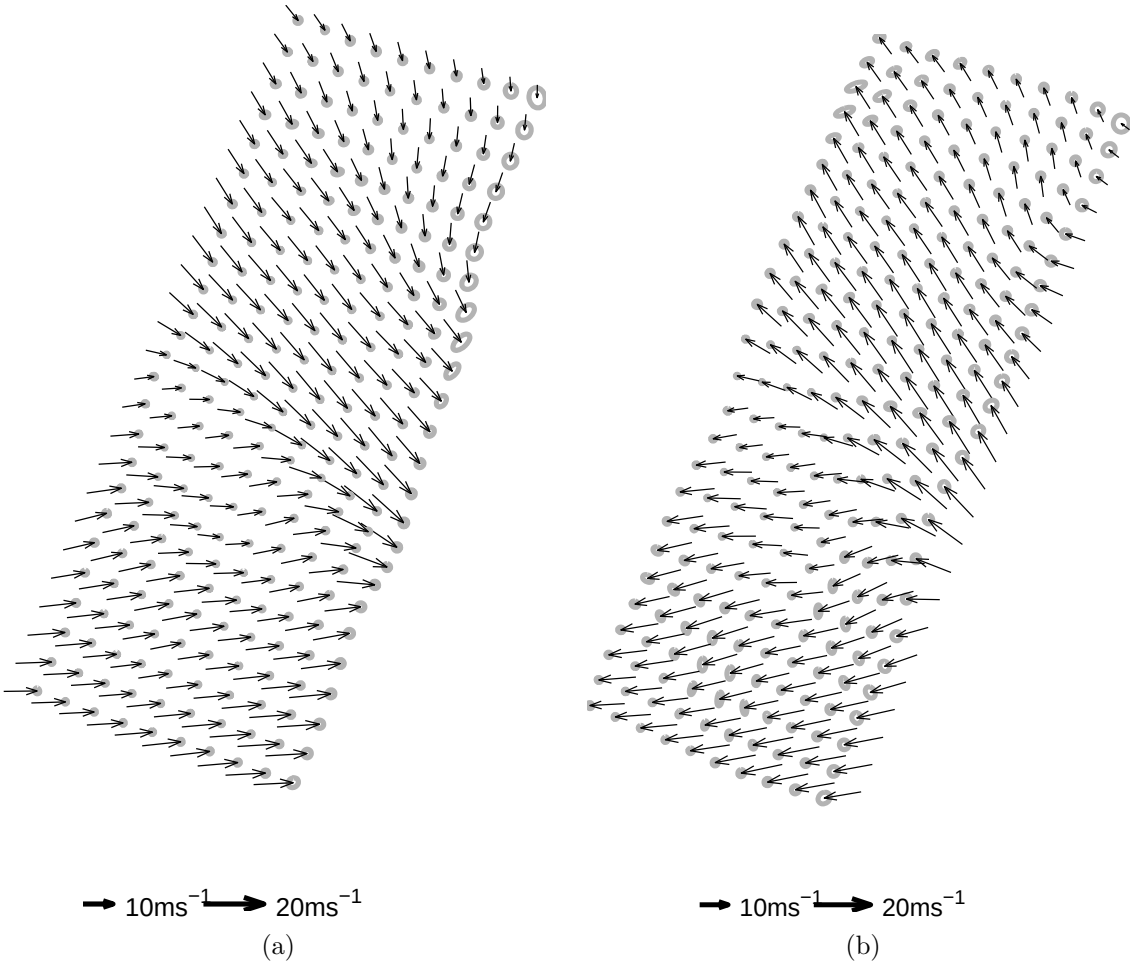

Fig. 2. Hybrid Monte Carlo samples using the inverse model from a North-West Atlantic wind £eld, on 10th June 1998, showing the two dominating modes. The black arrows de£ne the mean wind vectors, while the grey ellipses around the tips of the arrows de£ne the marginal covariance (plotted at 5 standard deviations to aid visualisation) at each observation.

The combination of the projection method, the sparsification technique and the expectation propagation algorithm constitutes the sequential Bayes algorithm used in the simulations shown later. Our experiments show that often the size of the basis vector set can be chosen to be a small fraction of the overall number of data without loosing too much predictive power.

In practice the algorithm requires us to compute the marginal likelihood $\tilde{Z}_{t+1}$, as in eq. (11). For some models and likelihoods this can be computed analytically (e.g. the inverse model, where the likelihood is a Gaussian mixture model), however for the non-linear forward model this is not analytically tractable. In this work we linearise the forward model, eq. (5), about the marginal posterior mean at time t, which leads to a Gaussian posterior. The linearisation means that it is important that the predicted marginal posterior distribution, $q_t(\boldsymbol{v}_{t+1})$ is reasonably close to a local maximum *a posteriori* probability solution, thus when using the forward model a non-zero mean prior is required.

5. Results and discussion

We will illustrate the methods in sections 4.1 and 4.2 on the retrieval of two example wind fields. First, we briefly evaluate the performance of the local forward and inverse models.

Table 1. A summary of the results comparing the forward and direct inverse approaches to local modelling.

	VRMSE (ms^{-1})	*Spd. std.* (ms^{-1})	*Dirn. std.* (°)	within 20° %	*Time* (s pattern^{-1})
Forward	2.7	1.7	16	53	1.05
Inverse	2.6	1.8	17	44	0.03

5.1. Forward models and direct inverse models

A test set of 60,000 independent scatterometer observations was processed using an inversion of the forward model, by finding the most probable wind vector given the scatterometer observation with a uniform prior over $\boldsymbol{v}$, and the direct inverse model. This allows us to assess the accuracy of the models with respect to the NWP observations, which is the closest thing we have to "ground truth". The results in Table 1 show that both models are very comparable in performance as might be expected. The forward model is considerably better at "guessing" the correct direction, Seeking the correct solution within 45°, that is in the correct quadrant, means the figures rise to 63% and 54% respectively, showing the inverse model still has skill in determining the correct from the ambiguous solution. The penalty to pay for using the forward model comes in the time taken to process one pattern: it is nearly two orders of magnitude slower than the direct inverse model.

5.2. Sampling and sequential Bayesian learning

The results of the hybrid Monte Carlo sampling from eq. (3) using a zero mean prior from the two dominating modes is shown in Fig. 2; (a) shows the "true" mode, while (b) illustrates the ambiguous mode. There were $10,000$ samples drawn from each mode, the results for the figures were computed on the last $5,000$ samples to account for burn-in and convergence was assessed using multiple chains. As expected with wind fields with moderately high wind speeds there is low posterior variance in wind speed and direction. Uncertainty at the edges is greater than that in the middle of the wind field, and this is due to the reduced impact of the spatial prior at the edges.

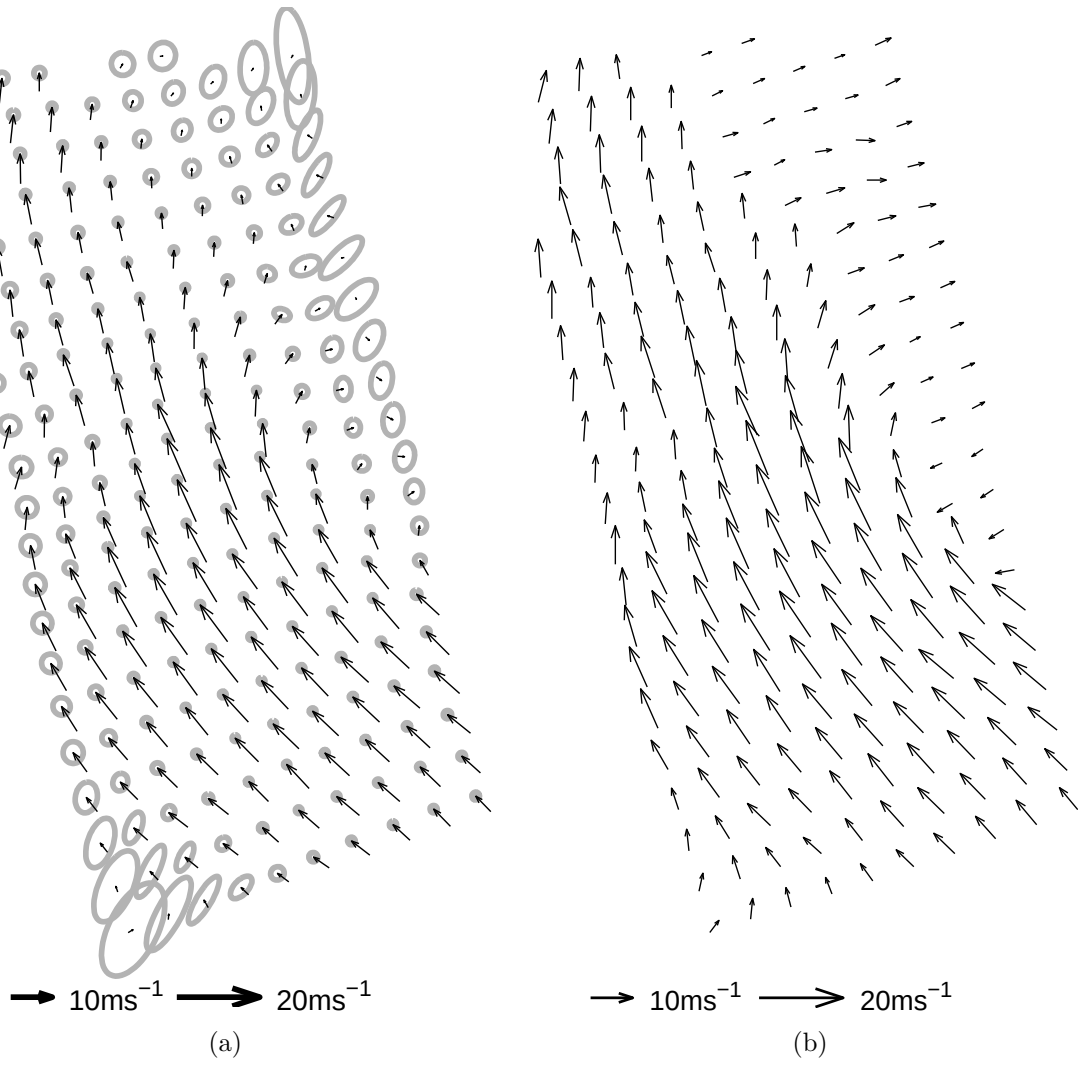

Fig. 3. (a) Hybrid Monte Carlo samples using the inverse model from a South Atlantic wind £eld, on 25th January 1999. (b) the corresponding UK Met. Of£ce wind vector estimates using the scatterometer data and NWP predicted winds.

Fig. 3 (a) shows the results of sampling from the most probable mode of the posterior distribution eq. (3). An interesting feature of this plot is the increased uncertainty in the local models where there is low wind speed. This correlates with the physical characteristics of the data collection system, where it is know that at low wind speeds wind direction is difficult to predict. This is evident by the larger ellipses in the upper right and lower left quadrant of the wind field. At low wind speeds, the two dominant solutions in local models are not completely separated, thus depending on the certainty in the neighbourhood, it is possible for wind vectors to swap into their ambiguous state. Also, this plot highlights how spatial consistency in the measurements increases certainty in the posterior distribution.

In order to assess the probability mass under the two dominating modes in both June and January wind fields $40,000$ samples were sub-sampled from a sample chain of $2,000,000$ points obtained using the mode jumping sampler described in Section 4.1. For each example the two maximum *a posteriori* probability modes were supplied to the sampler. After burn-in, each sampler remained in the dominate mode suggesting with a very high certainty that this is the correct mode. This can be explained by the factorising likelihood in eq.(3). The likelihood is a product of the local model probability distributions. If the local mass associated with the local wind vectors is 0.51 for the first mode and 0.49 for the second, then when these are combined in the likelihood, the total likelihood of the first mode would be $P(m_1) = \frac{(0.51)^N}{0.51^N + 0.49^N}$ where N is the number of wind vectors. Hence, when $N = 200$, $P(m_1) = 0.9997$, and the majority of the mass of the posterior is in the first mode although it is almost evenly distributed between each mode in the local model. In other experiments (Evans, 2001) we have shown that for smaller wind fields with $N = 100$ using the probability mass from sampling allows us to choose the correct wind field 73% of the time a small improvement over using maximum *a posteriori* probability.

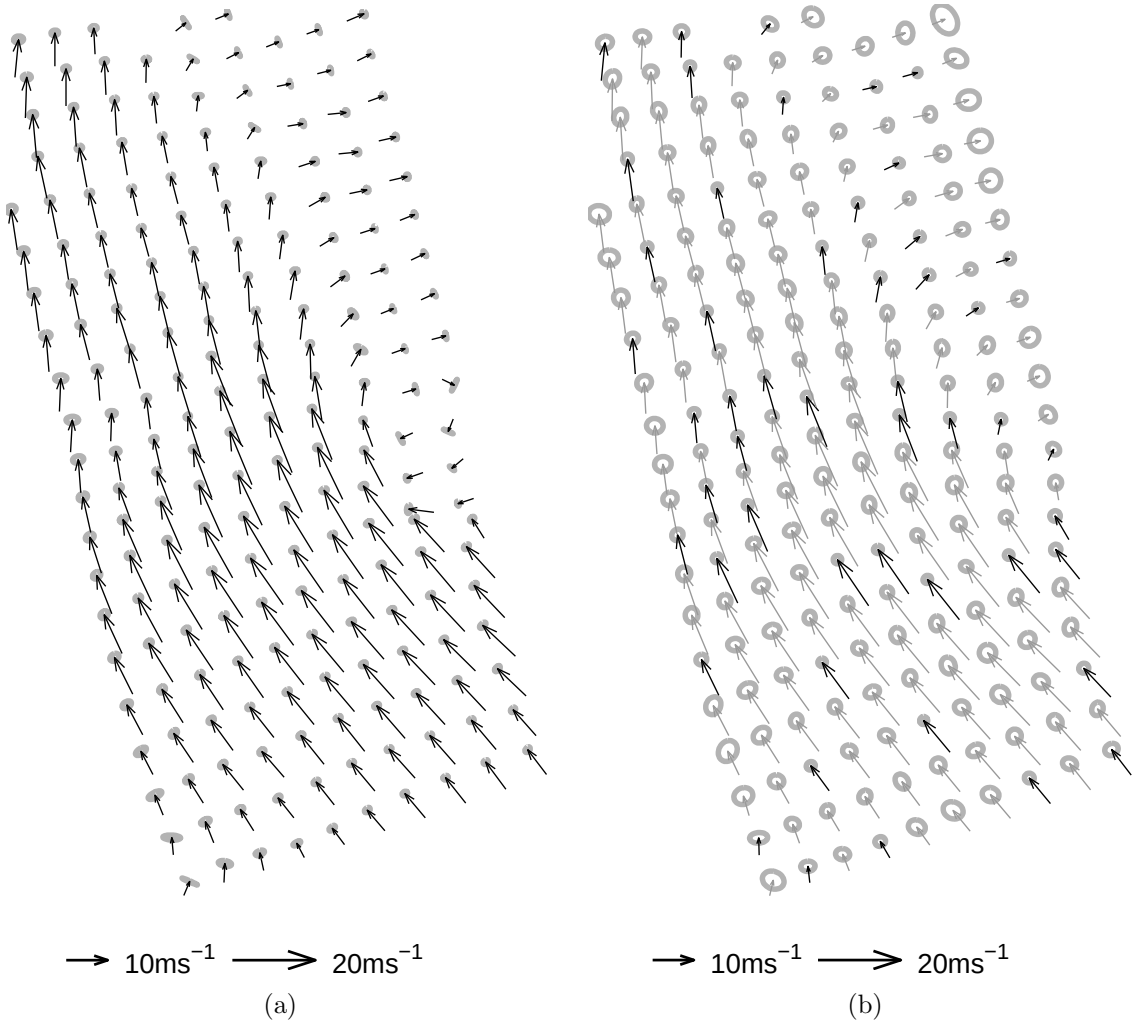

Fig. 4. The results of sampling from the posterior distribution eq. (1) for the wind £eld in Fig. 3 using the forward model and HMC sampling (a), and the sparse sequential Bayesian learning algorithm (b). In contrast to previous £gures these results use a non-zero mean Gaussian process prior, which represents the error statistics of the NWP forecast wind £eld, which forms the mean. In (b) only the (50) basis vectors are drawn in black, the remaining wind vectors are interpolated, shown in grey.

Fig. 4 shows the results of sampling using the forward model defined in Section 2.1 and using the Bayesian learning algorithm. It can be seen that the retrieved distributions are broadly similar, but that the Bayesian learning result (Fig. 4(b)) is smoother and has larger posterior uncertainty. This can be traced to the approximate nature of the sparse representation; we have run the experiment allowing more basis vectors to be used and the results tend to the sampling solution (Fig. 4(a)) as all points are included as basis vectors. In previous work (Csató et al., 2001) we have shown that sparse representations can well approximate the mean function, but that the covariance estimates degrade much more

quickly, as sparsity increases.

The key benefit of the Bayesian learning algorithm is the speed. To obtain the samples from the posterior distributions using both inverse and forward models took many hours of CPU time (the exact time depends on the number of samples and the size of the basis vector set), for just one wind field. This is impractical for operational use, where approximately 200 wind fields of the size shown are observed by the scatterometer each day. The Bayesian learning, on the other hand proceeds very quickly; 20 sweeps through the data using the expectation propagation method take less than two minutes of CPU time, on a 2 GHz Pentium machine. Although we do not show the results of applying Bayesian learning to the posterior distribution evaluated using the direct local inverse model, as the results are very similar to the forward model results, this is even faster since the integrals required for the update step in the algorithm can be computed analytically. When using the forward model it is necessary to linearise the model at the current state estimate to derive an expression for the integral, and this approximation means more sweeps through the data are required for the algorithm to converge, as well as each step being more computationally expensive.

6. Conclusions

The results have shown how local forward and direct inverse models have been constructed for scatterometer wind retrieval. We have described a Bayesian framework for field-wise retrieval of winds and have shown novel sampling and Bayesian learning approaches to the solution of this problem.

For the resulting wind fields to be used sensibly it is necessary to supply a mean prediction **and** the associated uncertainty. For the wind field retrieval problem this necessitates sampling or use of the Bayesian learning algorithm. The fact that the two methods give consistent results gives us confidence in the use of the Bayesian learning algorithm, which can use either direct inverse or forward models.

One drawback with using the forward models in wind field retrieval is that it is not possible to start with a zero mean prior: one either has to start with the NWP winds as the initial field, which results in a solution similar to Fig. 3(a) or use a non-zero mean prior as is done in Fig. 4. This is because the forward model is non-linear, thus both HMC sampling and the Bayesian learning algorithm use the Jacobian of the forward model, evaluated at the current estimate of the state. With a zero mean initial state this Jacobian is poorly defined and the implied linearization is a very poor approximation to the true model.

In practice the Bayesian learning algorithm allows us to treat large wind fields, much bigger than the ones we have dealt with in the examples shown, in a consistent framework which minimises the loss of information caused by the approximation, while allowing a fully probabilistic representation of the posterior distribution. While sampling does allow us to account for the non-Gaussian nature of the posterior distribution, all practical uses of the wind fields would only use the first two moments in any case, so this restriction to a Gaussian posterior is not too restrictive.

In future work we plan to use additional sources of information, such as cloud drift winds, to obtain completely independent estimates of the near surface wind field, by combining the multiple sources of information in the Bayesian framework proposed. This will be particularly important in the retrieval of wind fields when NWP forecasts are not available or desirable, for instance if the retrieved winds are to subsequently be used for estimating the state of the atmosphere in NWP data assimilation (where the impact of the NWP forecast could be very detrimental). In theory, the sequential Bayes algorithm allows us to compute a good approximation to the total probability of all observations, ie. the Bayesian evidence. This would give us a yardstick for comparing models with different hyperparameters.

References

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Bishop, C. M. and C. S. Qazaz (1997). Regression with input-dependent noise: A Bayesian treatment. In M. C. Mozer, M. I. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing*, Volume 9, pp. 347–353. MIT press.

Brooks, S. P. (1999). Markov Chain Monte Carlo and its Application. *The Statistician 47*, 69 – 100.

Bullen, R. J., D. Cornford, and I. T. Nabney (2003). Outlier detection in scatterometer data: Neural network approaches. *Neural Networks 16*, 419–426.

Cornford, D. (1998). Flexible Gaussian Process wind field models. Technical Report NCRG/98/017, Neural Computing Research Group, Aston University, Aston Triangle, Birmingham, UK. URL: `http://www.ncrg.aston.ac.uk/~cornfosd/`.

Cornford, D., I. T. Nabney, and G. Ramage (2001). Improved neural network scatterometer forward models. *Journal of Geophysical Research - Oceans 106*, 22331–22338.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. John Wiley & Sons.

Csató, L., D. Cornford, and M. Opper (2001). Online learning of wind-field models. In *International Conference on Artificial Neural Networks*, pp. 300–307.

Csató, L. and M. Opper (2002). Sparse on-line Gaussian Processes. *Neural Computation 14*(3), 641–669.

Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge: Cambridge University Press.

Dickinson, S. and R. A. Brown (1996). A study of near-surface winds in marine cyclones using multiple satellite sensors. *Journal of Applied Meteorology 35*, 769–781.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics Letters B 195*, 216–222.

Evans, D. J. (2001, August). *A pragmatic Bayesian approach to wind field retrieval*. Ph. D. thesis, Aston University, Aston Triangle, Birmingham. B4 7ET.

Haltiner, G. J. and R. T. Williams (1980). *Numerical Prediction and Dynamic Meteorology*. Chichester: John Wiley.

Handcock, M. S. and J. R. Wallis (1994). An approach to statistical spatio-temporal modelling of meteorological fields. *Journal of the American Statistical Association 89*, 368–378.

Hastings, W. K. (1970). Monte Carlo Sampling Methods using Markov Chains and Their Applications. *Biometrika 57*(1), 97 — 109.

Janssen, P. A. E. M., H. Wallbrink, C. J. Calkoen, D. van Halsema, W. A. Oost, and P. Snoeij (1998). Viers-1 scatterometer model. *Journal of Geophysical Research 103*, 7807–7831.

Kalman, R. and R. Bucy (1961). New results in linear filtering and prediction theory. *Trans ASME, Journal of Basic Engineering, Ser. D 83*, 95–108.

Levy, G. (1994). Southern-hemisphere low-level wind circulation statistics from the SeaSat scatterometer. *Annales Geophysicae - Atmospheres, Hydroshperes and Space Sciences 12*, 65–79.

Lorenc, A. C., R. S. Bell, S. J. Foreman, C. D. Hall, D. L. Harrison, M. W. Holt, D. Offiler, and S. G. Smith (1993). The use of ERS-1 products in operational meteorology. *Advances in Space Research 13*, 19–27.

Marinari, E. and G. Parisi (1992). Simulated Tempering: a New Monte Carlo Scheme. *European Physical Letters. 19*, 451 – 458.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics 21*(6), 1087 — 1092.

Mezard, M., G. Parisi, and M. Virasoro (1987). *Spin Glass Theory and Beyond*. Singapore: World Scientific.

Minka, T. P. (2000). *Expectation Propagation for Approximate Bayesian Inference*. Ph. D. thesis, Dep. of El. Eng. & Comp. Sci.; MIT, vismod.www.media.mit.edu/∼tpminka.

Morgan, N. and H. A. Boulard (1995). Neural networks for statistical recognition of continuous speech. *Proceedings of the IEEE 83*, 742–770.

Neal, R. M. (1995). Sampling from Multimodal Distributions using Tempered Transitions. *Statistics and Computing 6*, 353 – 366.

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

Opper, M. and D. Saad (Eds.) (2001). *Advanced Mean Field Methods: Theory and Practice*. The MIT Press.

Opper, M. and O. Winther (2001). Adaptive and self-averaging TAP mean field theory for probabilistic modeling. *Physical Review E 64*(056131), 1–14.

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

Stoffelen, A. (1998). *Scatterometry*. Ph. D. thesis, Universiteit Utrecht.

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

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

Tarantola, A. (1987). *Inverse Problem Theory*. London: Elsevier.

Tjelmeland, H. and B. K. Hegstad (2001). Mode Jumping Proposals in MCMC. *Scandinavian Journal of Statistics 28*, 205–223.

Ward, B. and S. Redfern (1999). A neural network model for predicting the bulk-skin temperature difference at the sea surface. *International Journal of Remote Sensing 20*, 3533–3548.