Estimation of Ecological Model Parameters by Implicit Sampling

B. Weir*, R. Miller, and Y. Spitz
College of Earth, Ocean, and Atmospheric Sciences
Oregon State University

* bweir@oce.orst.edu

(1) Motivation

We present a new methodology for data-based state and parameter estimation and results from its application to an ecological model. The method has a strong theoretical justification; it is appropriate when the model is nonlinear and for non-Gaussian distributions of the state conditioned on the observations; and it is able to estimate any number of parameters, including initial conditions and model error covariances. The ecological model describes the evolution of plankton, nutrients, and organic matter in the upper ocean. It depends on a set of parameters that determine the specific growth rates of the individual concentrations, their limiting due to crowding, and their interaction, whether it is consumption of one by another, or competition over a common resource. The true values of these parameters are rarely known precisely. Their estimation is essential to drawing accurate inferences about past, present, and future ecological states.

(2) Implicit estimation

The model is a discrete time stochastic process

$$\mathbf{X}_m = \mathbf{X}_{m-1} + \tau f(\mathbf{X}_{m-1}, \theta, t_{m-1}) + \sqrt{\tau} G(\mathbf{X}_{m-1}, \theta, t_{m-1}) \Delta \mathbf{W}_m.$$

At a subsequence of the model steps, we observe a noisy function of the state

$$\mathbf{Y}_n = h[\mathbf{X}_{m(n)}(\theta^*), \theta^*, t_n] + \sqrt{R}\mathbf{D}_n.$$

The unknown, true model parameters are the elements of the vector θ^*.

Begin with a set $\left\{\mathbf{x}_0^{(i)}\right\}$ of initial conditions for the particles (i.e., samples), which have the initial pdf $p(\mathbf{x}_0)$. For each particle, the implicit smoother does the following

1. Define the **cost function** $\mathcal{J}$ such that
$$p(\mathbf{x}_{0:m(k)}, \theta | \mathbf{y}_{1:k}) \propto p(x_0) \exp(-\mathcal{J}).$$
2. Minimize the cost function.
3. Sample a multivariate normal distribution about the minimum with the covariance matrix given by the inverse Hessian of the cost.
4. Weigh the particle by the ratio of the target pdf and proposal Gaussian.

The result is a set of weighted particles

$$\left\{\mathbf{x}_{0:m(k)}^{(i)}, \theta^{(i)}, w^{(i)}\right\} \sim p(\mathbf{x}_{0:m(k)}, \theta | \mathbf{y}_{1:k}).$$

(3) A simplified example

Consider the Lotka-Volterra equations for two species, a **predator** Q and **prey** P, with the grazing term in Michaelis-Menten form:

$$\frac{dP}{dt} = P(\theta_1 + \theta_2 P) + \theta_3 \frac{PQ}{1 + \theta_7 P} \quad (4a)$$
$$\frac{dQ}{dt} = Q(\theta_4 + \theta_5 Q) + \theta_6 \frac{PQ}{1 + \theta_7 P}. \quad (4b)$$

The equations have two types of invariant sets: a fixed point and a limit cycle. The fixed point is a center iff $\theta_2 = \theta_5 = 0$. Otherwise, it is a spiral. We take $a_5 = 0$, since the concentration of the prey limits the predator. In order to keep the solutions of equation (4) bounded, we constrain the sign of the state and parameters. A description and the units of each is given in the table below.

Table 1: The elements of the state and parameters. The sign of each variable precedes it.

		Description	Units
$(+)P$,	$(+)Q$	Concentration of prey, predator	# per sq. meter
$(+)\theta_1$,	$(-)\theta_4$	Specific growth rate of prey, predator	# per day
$(-)\theta_2$,	$(-)\theta_5$	Density dependence	(# per day) $\cdot$ (# per sq. meter)$^{-1}$
$(-)\theta_3$,	$(+)\theta_6$	Loss/growth due to grazing	(# per day) $\cdot$ (# per sq. meter)$^{-1}$
$(+)\theta_7$		Inverse half-saturation of grazing	(# per sq. meter)$^{-1}$

We use noisy measurements of a twin reference solution of the model with the true parameters to generate synthetic observations.

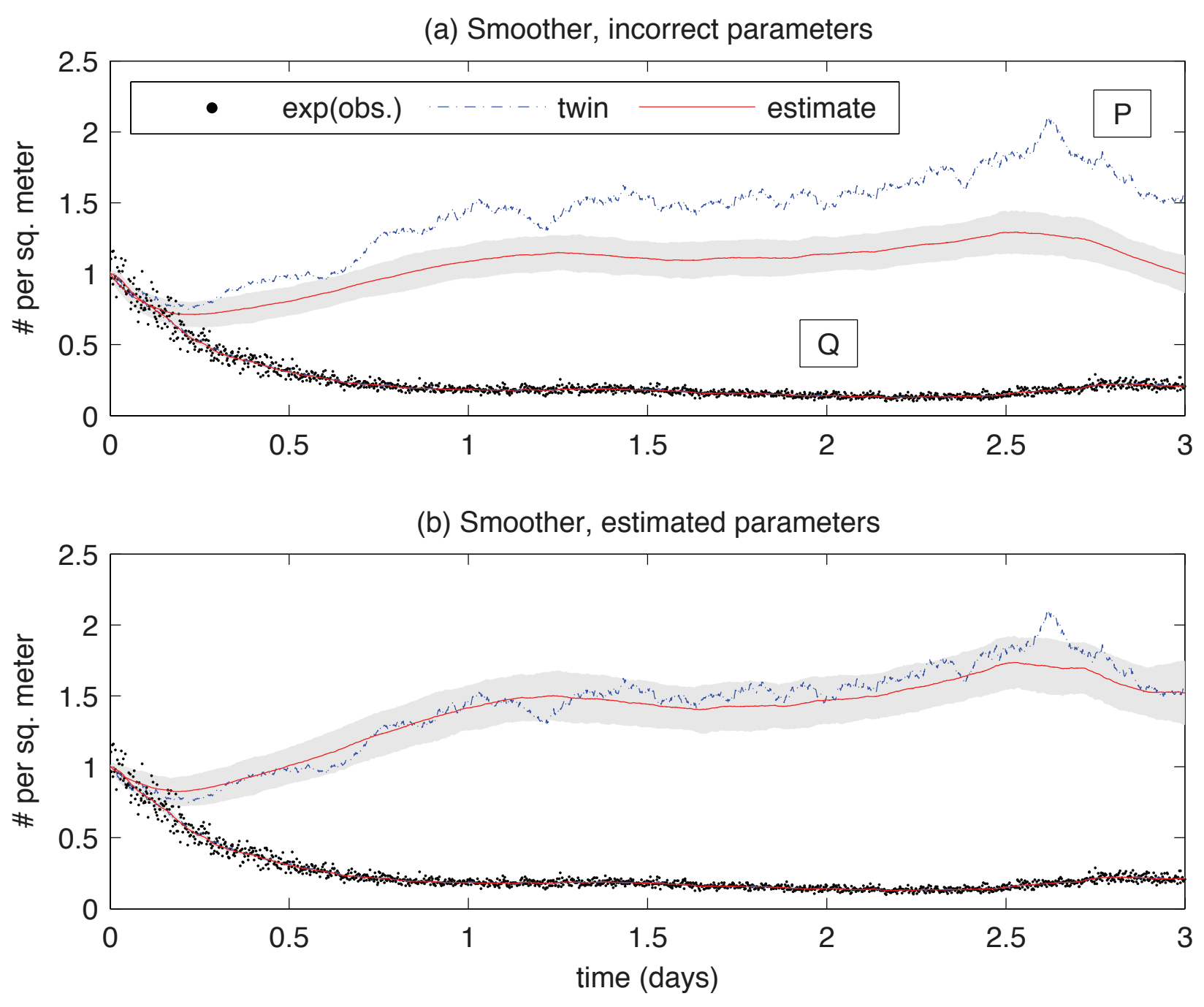


Fig. 1. Observations, reference path, and the state estimate computed with 240 particles when (a) θ_2 is held fixed at -4 and (b) θ_2 is estimated. Both estimates are computed using 240 particles. The shaded regions represent two standard deviations about these estimates. That the prey P in frame (a) is always outside this region demonstrates the failure of the model noise to account for the parameter error. This is corrected by estimating parameters in frame (b).

(4) A realistic biogeochemical model

The full model (Spitz et al., DSR-II, 2001) has 11 state variables and 49 parameters. Solar irradiance and mixed layer depths are taken as the monthly mean climatology.

Twin experiments, deterministic model

1. Monthly data gathered in situ given by the red boxes and arrows in Fig. 3 and the sum of the green words, which is particulate organic matter (POM). *The smoother with $O(100)$ particles produces parameter distributions comparable to Fig. 2a.*
2. Just chlorophyll-a, global coverage every $O(8)$ days with satellite-based instrumentation. The quadratic aproximation overpredicts the spread of the true cost (Fig. 4) and its eigenvectors do not line up with the correct directional dependence (Fig. 5). Gaussian importance sampling is thus inefficient. To correct, with each sample z update the Hessian such that

$$M_{n+1} = M_n + \epsilon_n [J(z) - K(z; H)],$$
$$H_{n+1} = \exp M_{n+1},$$

where $\epsilon_n = Cn^{-\alpha}$, J is the true cost, and K its quadratic approx.

Fig. 3. Flowchart of the annual nitrogen cycle. Thicker arrows indicate greater fluxes. The two exiting arrows represent sinking out of the mixed layer, while the entering arrow represents entrainment due to mixed layer deepening.

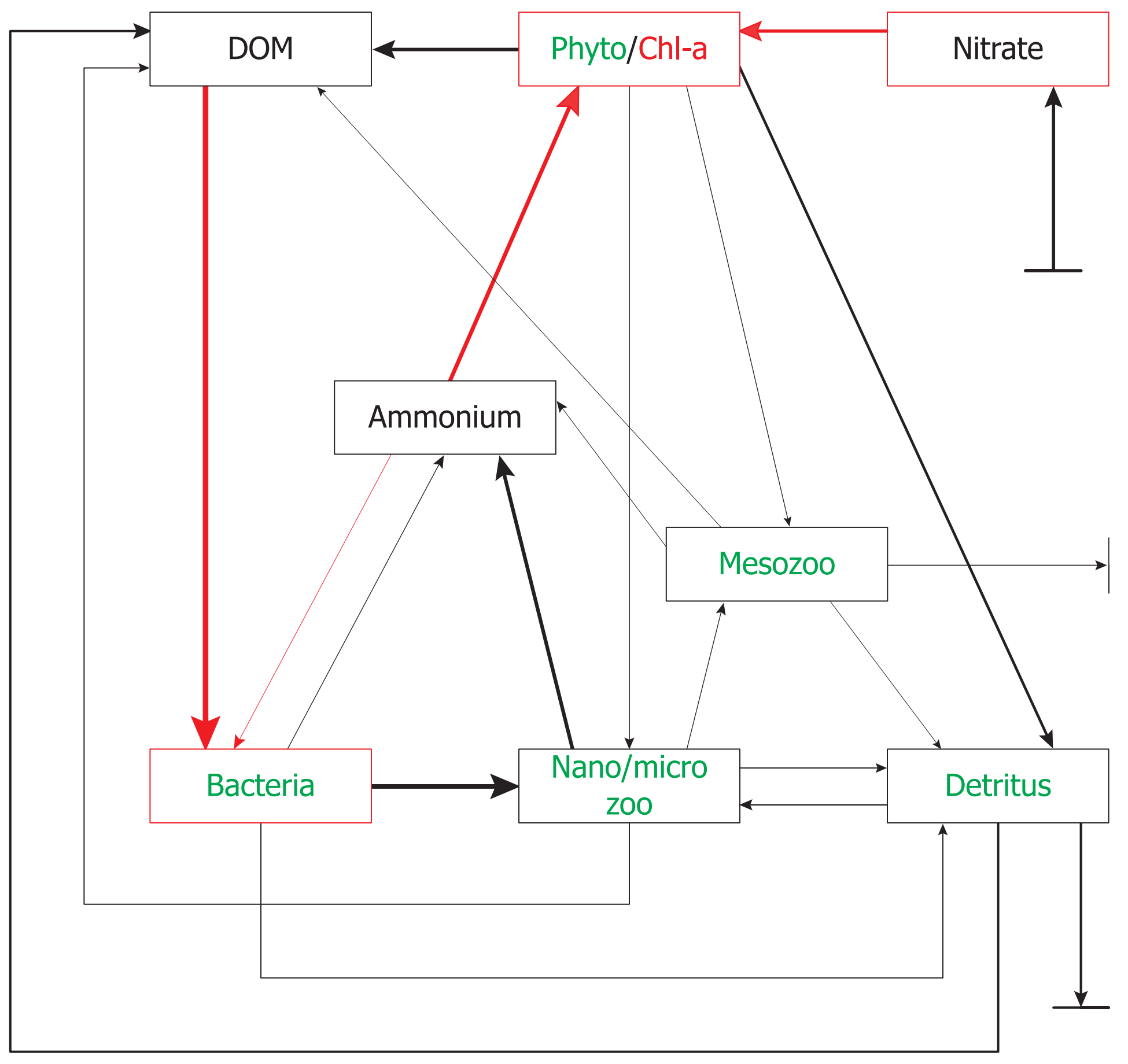


Fig. 4. The cost function and its quadratic approximation along a scaled eigendirection of the Hessian. The true cost diverges from the quadratic due to the importance of higher order terms.

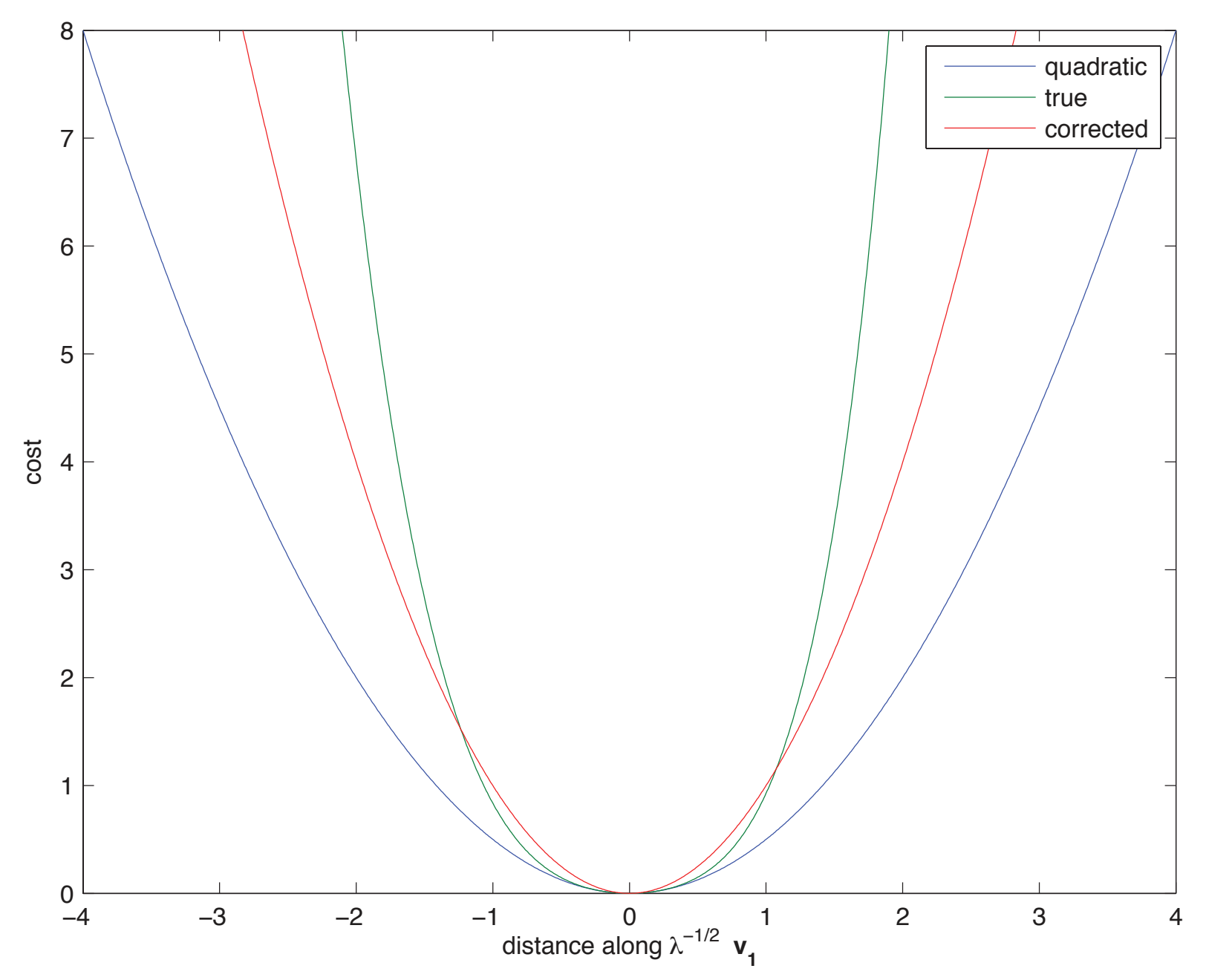
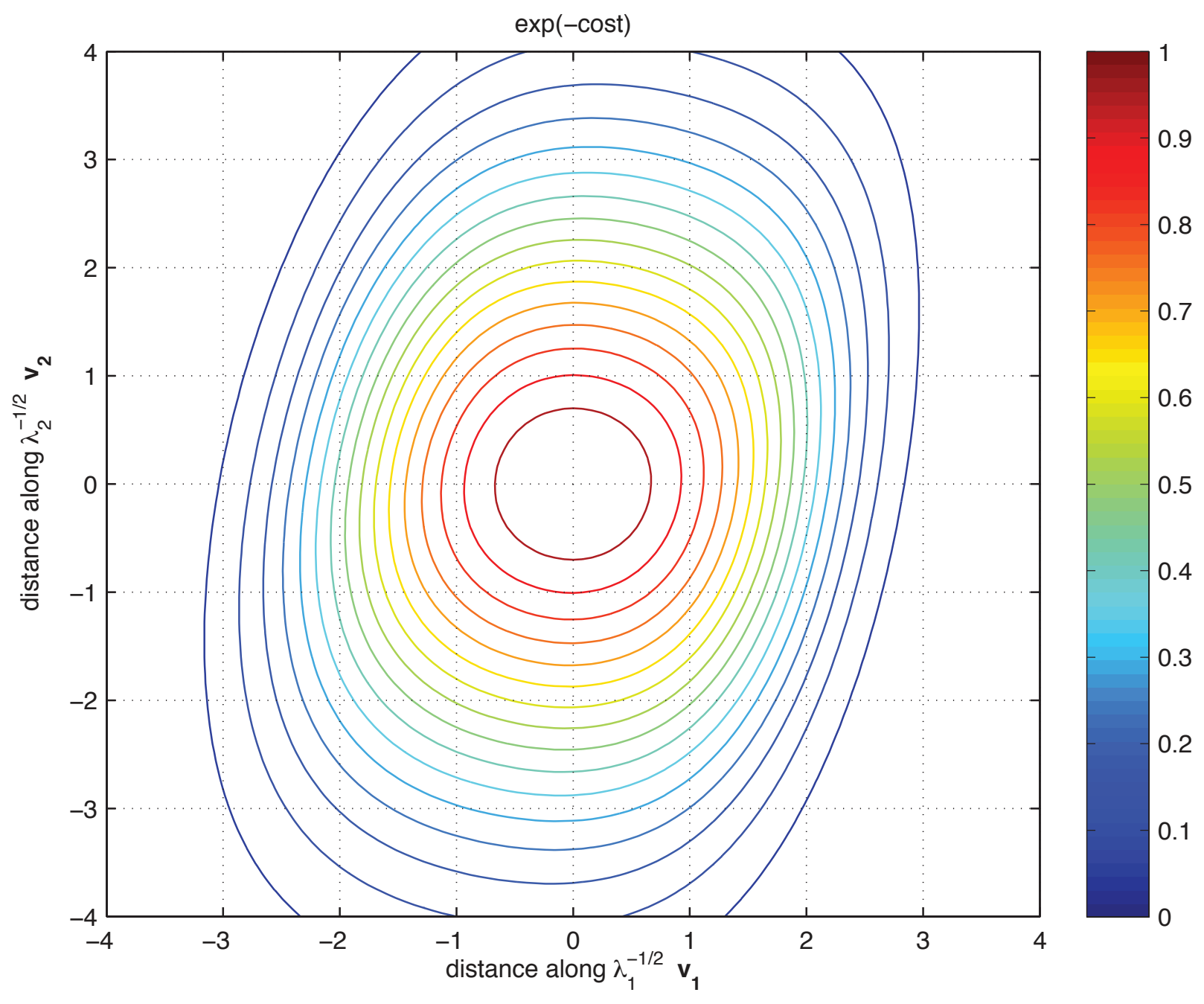


Fig. 5. Contours of the target density in two directions. The x and y axis correspond to eigendirections of the Hessian. Notice that the true cost varies in different directions.



(5) Conclusions

For a simplified example, the implicit smoother outperforms the estimates of SIR and EnKF. It is also able to make accurate inferences about a realistic biogeochemical model if there is sufficient data. When the data is chlorophyll only, the quadratic approximation of the cost severely overpredicts its spread. This leads to confidence intervals on the estimates that are unnecessarily large, and can cause approaches that rely on Gaussianty like EnKF to fail. We fix this deficiency by using a Robbins-Monro iteration to adaptively refine the Hessian each time we sample the proposal.

The authors thank E. Atkins, A. J. Chorin, and M. Morzfeld for their time spent in collaboration on this research. This work was supported by NSF grant #0934956.

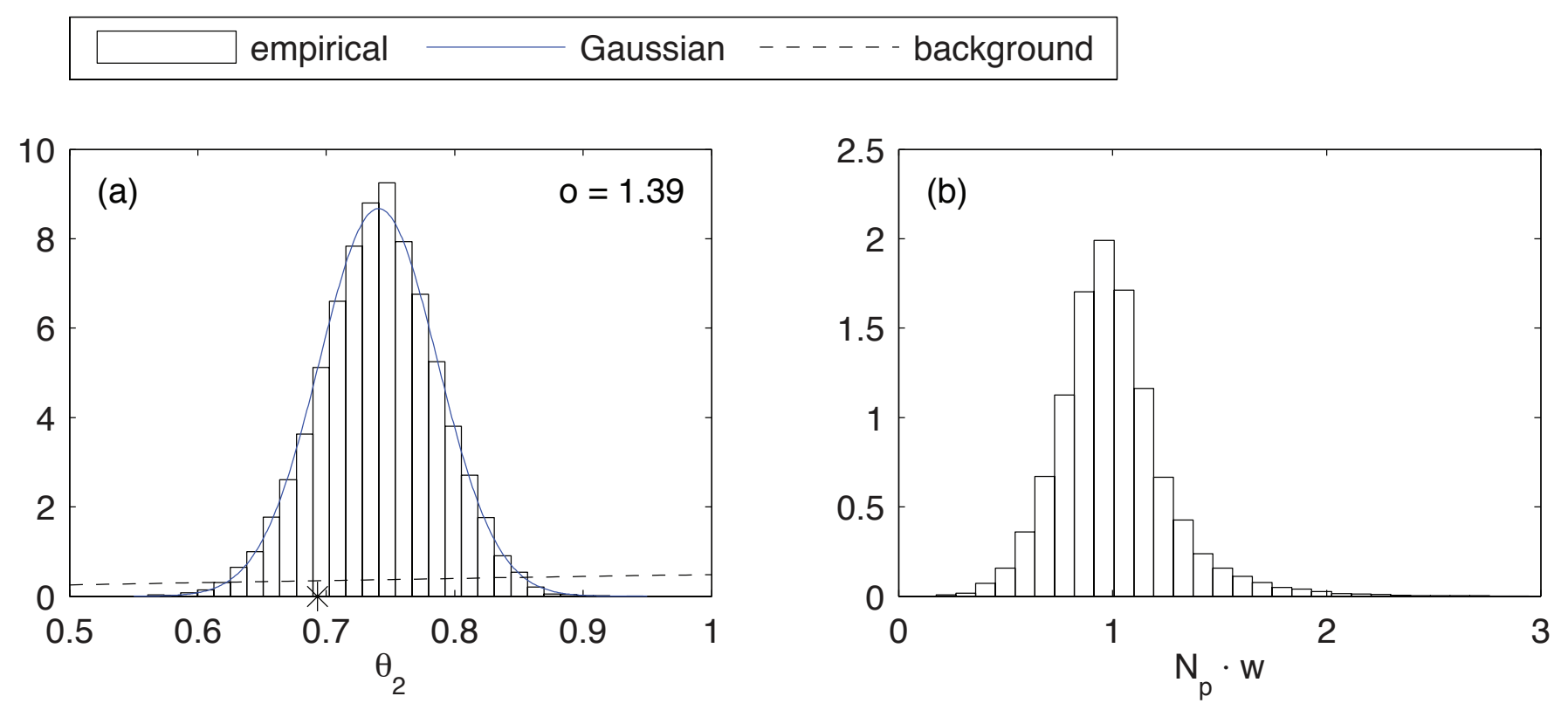
Oregon State University
College of Earth, Ocean, and Atmospheric Sciences
BERKELEY LAB
Lawrence Berkeley National Laboratory
NSF

(3 cont.) A simplified example

Table 1: Error statistics of 5000 trial estimates of log $-\theta_2$ for three different ensemble sizes and four different methods: the implicit smoother, the implicit filter, SIR, and EnKF. The numbers after the ± sign indicate the standard deviation of the statistics computed by resampling with replacement from the sample errors.

N_p	Mean	Median	RMS	IQR
		Implicit smoother		
24	-0.0351 ± 0.0133	-0.0007 ± 0.0028	0.9310 ± 0.1510	0.2540 ± 0.0050
240	-0.0005 ± 0.0088	0.0068 ± 0.0032	0.6137 ± 0.0292	0.2495 ± 0.0055
2400	-0.0079 ± 0.0119	0.0045 ± 0.0033	0.8210 ± 0.1332	0.2513 ± 0.0058
		SIR		
24	-0.3171 ± 0.0260	-0.0498 ± 0.0175	4.1652 ± 0.1556	3.0139 ± 0.0280
240	-0.1554 ± 0.0084	-0.0125 ± 0.0020	1.8577 ± 0.0628	0.5616 ± 0.0041
2400	-0.0057 ± 0.0038	0.0009 ± 0.0011	0.8514 ± 0.0252	0.2868 ± 0.0019
		EnKF		
24	-0.1277 ± 0.0116	-0.0017 ± 0.0039	1.8327 ± 0.0692	0.7731 ± 0.0077
240	0.0024 ± 0.0037	0.0025 ± 0.0011	0.8314 ± 0.0170	0.2988 ± 0.0019
2400	0.0136 ± 0.0036	0.0048 ± 0.0010	0.7979 ± 0.0178	0.2667 ± 0.0017

Fig. 2. Empirical representation of (a) the marginal $p(\theta_2 | \mathbf{y}_{1:k})$ of the target density and (b) the rank histogram of the weights of 24,000 samples. The Gaussian fit has the same mean and standard deviation as the samples.


empirical
Gaussian
background
(a)
o = 1.39
θ_2
(b)
$N_p \cdot w$