Do Box Inverse Models Work?

PETER C. MCINTOSH AND STEPHEN R. RINTOUL

Antarctic CRC, Hobart, Australia, and CSIRO Division of Oceanography, Hobart, Australia

Manuscript received 12 September 1995, in final form 24 June 1996

ABSTRACT

The performance of a box inverse model is tested using output from a near-eddy-resolving numerical model. Conservation equations are written in isopycnal layers for three properties: mass, heat, and salt anomaly. If the equations are free of error and the vertical exchange of properties between layers is negligible or known, the reference level velocity structure is quite accurately reproduced despite the underdetermined nature of the problem. If the interlayer fluxes of properties are not negligible and they are ignored, the solution for the reference level velocities is poor. If the interlayer fluxes of properties are included as additional unknowns in the inversion, they can be accurately estimated provided the column weights are chosen appropriately. Column weights that minimize the ratio of largest to smallest singular value (the "condition number") result in the best solutions for interfacial fluxes, and generally also for lateral fluxes. This choice of column weights also makes the inversion insensitive to data error: Inversions containing typical errors can be solved at full rank, obviating the need to estimate the rank. The choice of number of layers, and whether these layers are isopycnals or geopotentials, does not affect the accuracy of the inversion provided that interlayer fluxes are included as unknowns in the inversion. A reasonable estimate of solution accuracy is available by using the statistical approach to inverse problems, although this method can be sensitive to the choice of prior statistics.

Box inverse models do work, provided that they include interfacial fluxes as unknowns and that these are weighted appropriately. Such a model can successfully determine interfacial fluxes and, in some cases, horizontal fluxes. However, the model will not generally reproduce the detailed structure of the reference level velocities.

1. Introduction

Inverse methods have been used by oceanographers to determine the ocean circulation for more than 15 years. Although the term “inverse methods” includes a wide range of techniques, all inverse methods combine observations with dynamical information to estimate unknown features of the ocean circulation. The first application of inverse methods to physical oceanography was by Wunsch (1978), who addressed the classical problem of physical oceanography: determination of the reference level velocity, which must be added to the thermal wind velocity to give the absolute velocity.

The Wunsch formulation of the inverse problem has come to be known as the “box inverse” method. Hydrographic sections and land barriers are used to define closed volumes or “boxes” of ocean. Within each closed volume, the water column is divided into layers. Conservation of mass and other properties in each layer of each box results in a system of linear equations. Solution of these equations gives the velocity at the chosen reference level. In oceanographic applications, the system of equations is often underdetermined so that there are infinitely many solutions. A particular solution is usually selected by a least squares technique, which requires the reference level velocities to be close to an initial estimate (typically zero).

One of the strengths of the method is its flexibility: Essentially any prior knowledge can be incorporated as a constraint, and the method can be adapted to estimate a variety of unknown aspects of the circulation (e.g., diffusivities, Schlitzer 1987, 1988; Tziperman and Hecht 1988; Tziperman 1988). As a result, the box inverse method has become a commonly used tool for the analysis of oceanographic observations. However, the accuracy of a box inverse model applied to real data can be difficult to assess. A number of choices must be made when designing or solving a box inverse model (e.g., choice of unknowns, layer definition, constraints, weighting, rank). In the absence of knowledge of the true solution it is difficult to demonstrate that one choice is superior to another, so the inverse modeler is generally forced to try a number of possibilities and present a range of plausible results.

Numerical models provide an opportunity to test the inverse method in a situation where the true answer is known (e.g., Bigg 1986; Killworth and Bigg 1988). Provided that the numerical model used is a reasonable representation of the real ocean, it is likely that an inverse method that can successfully reproduce the absolute velocity field of the model will also be applicable.
to the real ocean. Similarly, if the inverse method fails to reproduce the model velocity field, one would have reservations about applying the method to real oceanographic observations. The goal of this paper is to provide answers to the questions faced by box inverse modelers working with oceanographic observations: How should I design a box inverse model, and how can I assess the accuracy of the solution?

Our box inversions are tested on output from the Fine Resolution Antarctic Model (FRAM Group 1991). FRAM is a marginally eddy-resolving model of the Southern Ocean, forced to be close to the Levitus (1982) climatology during the first six years of the integration. FRAM may be thought of as a state of the art description of the circulation of the Southern Ocean, in the sense that the model fields are consistent with both the observed climatology and nonlinear primitive equation dynamics.

Vertical sections of temperature and salinity, analogous to the observations one would obtain from a ship, are extracted from the FRAM model. A variety of inverse models are applied to the box defined by these sections and land barriers to assess the skill of the inversion in reproducing the known absolute velocity field. The general formulation of the box inverse model is presented in section 2, and aspects of FRAM relevant to the present study are described in section 3.

Many box inverse model studies assume that the exchange of properties between layers is unimportant. In section 4 we investigate how accurately the reference level velocities may be reproduced if this assumption is valid, both with and without random errors in the equations. We also assess what happens when this assumption is not valid.

In section 5, interlayer fluxes of properties are added to the inversion as unknowns. We investigate how accurately these may be determined in a variety of circumstances. The solution is shown to be sensitive to the weighting used and a method is developed to guide the optimal choice of weights.

One of the major issues in developing a box inverse model is the choice of layers. Most box inverse models use layers defined by approximate density surfaces. In section 6, we explore the consequences of using a very large number of isopycnal layers and also of using layers defined by geopotentials.

The main motivation for this study is to assess the accuracy of the box inverse method. The assessment is straightforward in this case because we are inverting synthetic data from a numerical model and we know the true velocity field. In an inversion applied to real data, the true velocity field will not be known, and less direct indicators of the accuracy of the inversion must be sought. In section 7 we assess the utility of the posterior covariance provided by one of the least squares methods. The implications of the study are discussed in section 8 and the main results are summarized in section 9.

2. Box model formulation

The standard box inverse model consists of conservation statements for mass and other tracers in a volume of ocean bounded by hydrographic sections or land in the horizontal, and isopycnal or similar layers in the vertical. In each subvolume, the sum of the horizontal and diapycnal fluxes of properties must nominally balance. The horizontal advective flux of a property consists of a baroclinic component, which may be estimated from the hydrographic data using the thermal wind relation, and a component due to the unknown reference level (or barotropic) velocity [see Wunsch (1978) for a more detailed description]. The horizontal diffusive flux can be included as an additional unknown, but is often ignored, especially when the spatial resolution of the data is high (e.g., Rintoul and Wunsch 1991).

We conserve mass, heat, and salt anomaly (no other tracers are carried in FRAM). Conservation equations for each property in each layer lead to a system of linear equations for the unknown barotropic velocities and diapycnal fluxes:

$$Ax = b.$$  \hspace{1cm} (1)

Vector $b$ contains the divergence due to the horizontal baroclinic property flux and the Ekman flux. It can be calculated from the observed density field using the thermal wind equation and from wind data using Ekman theory. Vector $x$ contains the unknown reference level velocities, and for some inversions, the unknown diapycnal fluxes. The elements of matrix $A$ consist of station pair layer areas multiplied by property concentrations. For a more detailed discussion of (1), see Wunsch (1978).

If interlayer fluxes are included in the inversion, $A$ also contains elements comprising horizontal interface areas multiplied by interface-averaged property concentrations. The unknown interfacial fluxes include both advective and diffusive contributions; we combine the diapycnal advective and diffusive fluxes into a net depth-dependent interfacial flux for each property. With $A$ defined as above, the interfacial flux unknowns have units of velocity and will be referred to as effective interfacial velocities. For mass, this will be the advective velocity. For other tracers, the effective interfacial velocity is the velocity that, when multiplied by the interface area and mean property concentration, gives the total flux and hence will contain contributions from both advection and diffusion. We will often refer to the diapycnal unknowns as fluxes rather than velocities to remind readers that, except for mass, the effective interfacial velocity does not represent a true diapycnal velocity. Vector $x$ consists of the unknown reference level velocities $v$ at each station pair and the unknown interfacial “fluxes” for mass, heat, and salt anomaly ($w_m$, $w_q$, and $w_s$, respectively):

$$x = [vw_mvw_qvw_s]^T.$$  \hspace{1cm} (2)

Our approach has two drawbacks: First, each new
property adds as many new unknowns as equations, so that conserving additional properties does not increase the determinacy of the problem. Second, the net interfacial flux appears in the conservation equation in advective form (velocity times concentration). Assuming that diffusion is parameterized in the usual Fickian manner, the diapycnal diffusion would generally be represented as a diffusivity multiplied by a diapycnal gradient. However, the representation of diapycnal fluxes in models such as these is not quite so simple. Because we can estimate the lateral fluxes only through the perimeter of each box, we can determine only the net interfacial flux over the area of each interface. While it is straightforward to represent the net flux as the sum of a mean diapycnal advection and a diffusivity multiplied by the mean diapycnal gradient of each property in the inverse model, an important term is left out: the flux due to correlations between variations of diapycnal velocity and concentrations on each interface. In FRAM this “Reynolds” term can be large relative to the explicit model vertical diffusivity. The interfacial flux question is further complicated by the diapycnal flux resulting from horizontal diffusion across steeply sloping isopycnals, which can also be significant in FRAM. Given these difficulties, we have taken the simplest approach, which is to represent the combined result of all processes responsible for carrying properties across interfaces as a single net interfacial flux (or more precisely, an effective interfacial velocity times the mean concentration on the interface). A more complete discussion of vertical and diapycnal processes in FRAM, and their representation in box inverse models, is in preparation.

We have used the model velocity shear in all inversions. The only significant departures from geostrophy in the FRAM data used here are in the Ekman layer. Use of the model velocity shear is essentially an easy way to include the correct Ekman flux on the right-hand side of (1). In an inversion using real data, the Ekman flux would need to be estimated from wind data. It is not our purpose to study this problem here.

Wunsch (1978) and others have noted the importance of scaling the system (1) properly. Row weighting involves weighting each equation by premultiplying \( \mathbf{A} \) and \( \mathbf{b} \) by a matrix \( \mathbf{W}_c \). The square of this matrix, \( \mathbf{W}_c^2 \), is formally the inverse of the equation error covariance matrix. In the absence of prior estimates of the full covariance matrix, \( \mathbf{W}_c \) is typically taken to be diagonal. Here we weight each equation by the reciprocal of the typical property value in each layer. We consider this a reasonable approximation to weighting each equation by the reciprocal of its standard deviation, which is desirable if the statistical interpretation of the inverse problem is adopted (Tarantola 1987). We note that if the system (1) is solved at full rank, then row weighting makes no difference to the solution obtained. It is only when residuals are allowed in the equations that the weighting of these residuals as part of a penalty function becomes important.

Column weighting involves postmultiplying \( \mathbf{A} \) by a matrix \( \mathbf{W}_r \) and premultiplying \( \mathbf{x} \) by the inverse of \( \mathbf{W}_r \). The square of this matrix, \( \mathbf{W}_r^2 \), is formally the prior solution error covariance matrix. Again, this matrix is usually chosen to be diagonal, with the size of the diagonal elements chosen to reflect the relative magnitude of the elements of \( \mathbf{x} \). An objective method for determining the optimal choice of column weights is developed in section 5.

The weighted system of equations is solved using the singular value decomposition (SVD) (Lanczos 1961). The SVD solution simultaneously minimizes the size of the solution \( \mathbf{x} \) and the size of the residuals \( \mathbf{A} \mathbf{x} - \mathbf{b} \) in a least squares manner. The SVD solution is thus the “simplest” solution in the sense that it introduces the minimum correction to the initial model required to satisfy the constraints. In the presence of errors in the model or the data, the number of singular values included in the solution will generally have to be truncated at something less than full rank to avoid amplification of errors. The performance of the inverse method in the presence of errors is a major focus of the experiments described in sections 4 and 5.

Because the SVD solution penalizes the size of the correction to the initial model, the solution to a box inversion can be sensitive to the choice of initial state. In the experiments discussed in this paper, the initial state is chosen to be zero velocity on one of the surfaces, or at the sea floor. The interfacial fluxes are initially set to zero in those experiments in which they are carried as unknowns. We choose the depth of the initial level of no motion as an oceanographer working with data from the real ocean would do: to be consistent with water mass properties and, if available, direct velocity measurements. For the Southern Ocean box, direct velocity measurements suggest that a deep reference level is appropriate (Nowlin and Klinck 1986). In the Tasman Sea, analysis of water masses suggests a middepth reference level is likely (e.g., Wyrtki 1961; Warren 1973). Some experiments exploring the sensitivity of the results to the choice of initial reference level are discussed below.

3. Validation data

The Fine Resolution Antarctic Model provides hydrographic and velocity “data” that are dynamically consistent in the sense of primitive equation physics. In addition, the FRAM temperature and salinity distributions are close to the Levitus climatology because of the robust diagnostic forcing used. In this sense, the FRAM ocean can be considered a state of the art description of the ocean circulation around Antarctica. Using FRAM output we can diagnose all of the terms in \( \mathbf{A} \), \( \mathbf{x} \), and \( \mathbf{b} \), and so rigorously test a method designed to estimate \( \mathbf{x} \).
We examine two regions: a Southern Ocean box bounded by north–south sections extending from Australia to Antarctica at 117° and 147°E, and a smaller Tasman Sea box bounded by sections extending from Australia to New Zealand at 36° and 43°S (Cook Strait and Bass Strait are closed in FRAM). The two boxes are shown in Fig. 1. The horizontal resolution of the FRAM model is 0.25° in the north–south direction and 0.5° in the east–west direction. This results in 224 and 92 unknown reference level velocities to be determined for the Southern Ocean and Tasman Sea boxes, respectively.

Layers are defined by locally referenced isopycnals, also known as neutral surfaces (McDougall 1987), and are computed as described by Jackett and McDougall (1997). Most of the experiments use 21 layers, defined somewhat arbitrarily to resolve the water masses seen in the temperature and salinity fields. Experiments with many layers, and with layers defined by geopotential surfaces, are discussed in section 6.

We use a single time step from the FRAM integration at the end of year 6 (Webb et al. 1991), thus ignoring issues related to the asynopticity of real hydrographic data. This is the last time step at which FRAM is being relaxed to the Levitus climatology. The relaxation acts as a significant source/sink term of heat and salt throughout the ocean volume. Because such a term is missing from real data, we treat this term as known and include it on the right-hand side of (1). We also diagnose the tracer time-dependent (storage) term and likewise include it in $b$. (In practice, only the sum of the relaxation and storage terms can be diagnosed from a single model snapshot, not their individual contributions.) The Ekman flux across each section is also taken as known and included in $b$. The temporal evolution of isopycnals cannot be calculated from a single snapshot and is ignored; hence, diapycnal fluxes are in error by the vertical advection rate of isopycnal surfaces. Other than the restoring term in the surface layer of FRAM (which we have included), there is no explicit surface heat or freshwater flux. As a consequence, there are no additional source or sink terms due to surface fluxes when layers outcrop at the sea surface (as many do in the Southern Ocean).

We calculate the advective and diffusive fluxes across the boundaries of each layer, together with the residual terms (relaxation plus time-dependent), and confirm that (1) holds exactly. This calculation in layers is not trivial because FRAM is a level model: The numerics used for the diagnosis must be consistent with the numerics of FRAM. In particular, we have found it necessary to approximate neutral surfaces by a series of steps, with vertical segments occurring at the edge of model tracer boxes and horizontal segments at the depth of the neutral surface as determined from data at model tracer points. The storage plus relaxation terms integrated over each layer are a significant fraction of the net property divergence due to the baroclinic and Ekman flow: typically 23% for heat, 30% for salt anomaly in the Southern Ocean, and 22% for heat, 57% for salt anomaly in the Tasman Sea.

As mentioned earlier, we write conservation equations for salt anomaly, rather than salinity. McDougall (1991) discusses the importance of using anomalies of
tracers, such as salinity, which have small variations about a large mean value. We conducted three preliminary experiments in which we used salinity — $S$, $S = S - 35$ and $S = S$, where the mean $S$ was taken over all sections and layers. The mean was typically within 0.5 of 35 psu. The singular values of $A$ were calculated for the Southern Ocean box for each case. The ratio of largest to smallest singular value (the condition number, a measure of the sensitivity of the inversion to errors) is increased by two orders of magnitude when $S$ is used compared to when $S = 35$ or $S = S$ is used. We conclude that use of a salinity anomaly is highly desirable but that it is not important that the reference salinity exactly equal the mean value of salinity. In the remainder of the paper we use salt anomaly defined as $S' = S - 35$.

The variation of temperature around its mean value is comparable to the mean value itself, and defining a temperature anomaly gives no significant improvement in the condition number.

4. Inverting for reference level velocities alone

a. Error-free inversions

We first address the question: How accurately can the reference level velocities be estimated under ideal conditions? We assume that all other quantities are known perfectly: The exact velocity shear is used, the known interfacial fluxes are added to $b$, and the (small) horizontal diffusive fluxes are also added to $b$. Hence, if $x$ equals the true reference level velocities, (1) is satisfied exactly. However, since the problem is underdetermined for both test regions (224 unknowns and 63 equations for the Southern Ocean box, 92 unknowns and 45 equations for the Tasman Sea box) there will be an infinite number of vectors $x$ satisfying (1) exactly. The particular solution selected by the full-rank SVD is the one which has the minimum sum of squared elements. This experiment tests whether the philosophy underlying the least squares solution — find the minimal correction to the initial state, required to satisfy the constraints — is appropriate for determining the circulation.

With a reference level at the bottom, the FRAM model velocity shear gives a transport of $211$ Sv ($Sv = 10^6 m^3 s^{-1}$) across the $117^\circ$E section, and $170$ Sv across $143^\circ$E, for a net convergence of $41$ Sv in the Southern Ocean box. The true model transport across $117^\circ$ and $143^\circ$E is $195$ Sv. For the Tasman Sea, the FRAM model transport relative to the bottom is $-7$ Sv at $43^\circ$S and $-9$ Sv at $36^\circ$S, for a net convergence of $2$ Sv, while the true flux across both sections is $-18$ Sv (negative values indicate southward flow). With an initial reference level at about $2800$ m, the net transport is $-7.4$ Sv at $43^\circ$S and $-6.7$ Sv at $36^\circ$S.

True and inverse solution velocities at the reference level for the Southern Ocean box are shown in Fig. 2, while those for the Tasman Sea are shown in Fig. 3. The initial reference level in each case was chosen to be the sea floor. The solutions in both regions are very close to the true velocity. Even features with relatively small spatial scale are well reproduced. The mass, heat, and salt anomaly fluxes through each section are reproduced to within 2% of the true values in the Southern Ocean and to within 4% in the Tasman Sea. An initial reference level at about $2500$ m in the Tasman Sea resulted in fluxes that were accurate to 1% or better.

We conclude that in the absence of error, and when the interfacial fluxes are known, the inverse method can reproduce accurately the absolute velocity field and horizontal fluxes of properties, even when the problem is significantly underdetermined. This is true even when the initial state is far from the true velocity field (e.g.,
an initial imbalance of 41 Sv in the Southern Ocean box, or an initial underestimate of the flow across the Tasman Sea sections by a factor of 2).

b. Inversions with errors

No inverse problem is ever free of error. One source of error is inadequacy of physics in the box model equations. The experiment discussed below in which interfacial fluxes are ignored is an example of an inverse model that is missing an important physical process. Error may also arise from inaccuracy in the data, such as errors in thermal wind due to measurement errors in temperature, salinity, or position; errors in wind stress used to calculate Ekman fluxes; discretization error; errors in surface heat or freshwater fluxes; and extrapolation error below the deepest common depth of a pair of hydrographic casts.

The behavior of the inverse solution in the presence of error is largely determined by the ratio of the largest to smallest singular value kept in the inversion. This diagnostic is called the condition number of the problem and indicates the amount by which relative errors in \( b \) might be amplified to become relative errors in \( x \) (Tarantola 1987). For error-free inversions, any condition number is acceptable. When interfacial fluxes are not included as unknowns, both boxes have condition numbers of around 1 \( \times \) 10^5, indicating that these inversions will be sensitive to even small errors in the data.

1) Inversions with data errors

We now look at what happens when the interfacial fluxes are considered known, as in the previous experiments, but other factors lead to errors in the equations. The size of data errors arising from the sources described above can be difficult to estimate and will depend on the particular dataset used. To evaluate the performance of the inverse method when applied to "noisy" data, we added 10% random errors to \( b \). We believe that errors in the geostrophic divergence calculated from real hydrographic data are likely to be smaller than this so that 10% errors pose a reasonably stringent test for the inversion. All experiments with random errors assume normally distributed errors with a standard deviation specified as a percentage of the typical size of elements of \( b \). The experiments are conducted over 100 realizations of the errors, and solution errors quoted are mean absolute values. No qualitative difference is obtained by using solution maximum errors, although the error level is higher.

With 10% random errors, a good solution is obtained for the Southern Ocean box for all ranks less than or equal to 33 (out of 63). (We define a "good" solution to be one having less than 10% mean absolute errors in mass, heat, and salt fluxes, and an rms reference level velocity no larger than the rms of the true reference level velocity. The second condition guards against solutions where the section fluxes are acceptable but the reference level velocities are much larger and more variable than might reasonably be expected.) For the Tasman Sea, no solution has flux errors less than 10% for mass, heat, and salt, but many solutions from rank 9 to 30 (out of 45) would be considered acceptable at the 20% flux error level. A discussion of the relatively poor results in the Tasman Sea relative to the Southern Ocean is deferred to section 5.

We conclude from the experiments with 10% random errors (and known interfacial fluxes) that the inversion is somewhat sensitive to error, as expected given the size of the condition number. To avoid amplifying errors, it is necessary to truncate the number of singular values used; however, choosing the appropriate truncation, or rank, can be difficult when inverting real data. As discussed in section 5, addition of interfacial flux unknowns improves the conditioning of the problem sufficiently that the difficult issue of determining the rank can be avoided.

2) Inversions with model errors: ignoring interfacial fluxes

As another example of an inversion in the presence of model error, we consider the case where the interfacial fluxes are neither considered known nor included explicitly as unknowns. The matrix \( A \) contains only areas and properties from the vertical sections comprising the side walls of the box, while the vector of unknowns, \( x \), contains only reference level velocities. Such a model was common in the early applications of inverse methods to oceanographic problems (e.g., Wunsch 1978; Wunsch et al. 1983; Roemmich and McCallister 1989). In the case of the Southern Ocean, this amounts to adding an error to \( b \) of around 70%. It is even worse for the Tasman Sea; the error is about 90%. In other words, the dominant contribution to the net divergence in the layers is the diapycnal divergence. These experiments can be thought of as exploring the consequence of leaving out an important physical process from the inverse model.

The mass, heat, and salt anomaly flux errors (averaged over both sections) are shown as a function of rank in Fig. 4 for the Southern Ocean box. Beyond a rank of about 8 the property fluxes across the sections begin to diverge rapidly from the true solution. In fact, the best flux estimates are obtained at low rank (rank \( \approx 3 \)). The reference velocities at rank = 1 are nearly uniform across each section, but the small uniform correction is sufficient to reduce the initial net mass divergence from 41 to 1 Sv. The mass imbalance in individual layers is 1–2 Sv for rank \( \approx 3 \).

As the rank is increased, the system introduces more structure to the reference level velocities in an effort to satisfy the conservation constraints in individual layers. In the FRAM model, much of the geostrophic mass divergence in layers is balanced by interfacial fluxes.
When interfacial fluxes are left out of the model physics, large noisy reference level velocities are required to reduce the residuals in the conservation equations. As a result, the net flux across each section is not well reproduced.

Some studies have calculated the implied interfacial velocities in similar inversions by integrating the layer residuals vertically from the bottom, rather than including the interfacial fluxes explicitly (e.g., Wunsch et al. 1983; Roemmich and McCallister 1989). However, inspection of the residuals in the present inversions suggests that the interfacial fluxes calculated in this way are likely to be a poor estimate of the true interfacial fluxes. The weighted least squares solution obtained by the SVD tends to leave residuals of roughly equal magnitude in each layer. At least in this example, integrating the residuals will not produce an accurate estimate of the interfacial fluxes.

Figure 5 shows mass, heat, and salt anomaly flux errors as a function of rank for the Tasman Sea box. The fluxes are poorly estimated at all ranks, with perhaps the exception of a rank near 6 or 25, where the curves happen to cross the axis. Note that, as for the Southern Ocean box, the low rank solutions in the Tasman Sea case conserve total mass. Again, the reasons for the relative difficulty of producing accurate flux estimates in the Tasman Sea are discussed in section 5.

We conclude that for both inversions, the interfacial fluxes are large enough and the conditioning of the inverse poor enough that it is impossible to obtain useful information concerning horizontal property fluxes if the interfacial fluxes are ignored. In addition, estimating the interfacial fluxes implicitly by vertically integrating the layer residuals is unlikely to produce good results.

5. Inverting for interfacial fluxes and reference level velocities

a. Error-free inversions

Interfacial fluxes may be included as unknowns in the inversion in a number of ways. As discussed in section 2, we choose the least restrictive way by defining a different net interfacial flux for each property at each interface. This flux contains both advective and diffusive components (except for mass). We begin by discussing error-free inversions.

For the Southern Ocean box, there are 21 layers defined by 20 surfaces plus the ocean surface and bottom. Hence, there are 20 unknown fluxes for each property, assuming no flux through either the surface (because of the relaxation) or through the bottom, for a total of 60 unknown fluxes to add to the 224 reference level unknowns. The number of equations remains at 63. In practice the elements of the extra columns in $A$ are comprised of surface areas times interface-average property values, calculated exactly from FRAM, so that the extra unknowns have units of velocity.

In the inversions discussed so far, we have used uniform weighting of the unknown reference level velocities. However, now that $x$ contains both reference level velocities and an effective diapycnal velocity (flux divided by area and property), it is necessary to choose a column scaling that trades off the size of the reference level and effective diapycnal velocities when the inversion minimizes the size of $x$.

The simplest procedure is to vary the relative weight and examine the effect on the solution. The most obvious effect is on the condition number of the problem.
For the Southern Ocean box, the condition number is plotted against the weight given to the interfacial fluxes relative to the weight given to the reference level velocities (Fig. 6). We see that for small values of the relative weight the condition number tends to the large values encountered in the experiments with no interfacial flux unknowns. That is, the weighting given to the interfacial velocities is so small that they effectively take no part in determining the solution. At the other extreme, the reference level velocities are heavily penalized, and all the information is pushed into determining the interfacial fluxes. In this case, there are effectively more equations than unknowns and the problem becomes nearly evenly determined. Hence, only the first 60 singular values are nonzero, and the condition number becomes large. At intermediate weights, the condition number is between 10 and 100, indicating that an inversion would be well conditioned and tolerant to some error. The Tasman Sea inversion exhibits a similar behavior.

We assess the performance of the inverse solution as a function of the column weighting in the case where there are no errors in the equations. For the Southern Ocean box, we find excellent solutions for all relative column weights above a value of about $1 \times 10^{-4}$. All these solutions have horizontal flux errors of less than 4%, and the correlation coefficients between the inverse solution interfacial fluxes and the true fluxes calculated from FRAM are around 0.9. However, the reference level velocities contain very little detail, even though the fluxes are well reproduced. The resolution matrices show that the equations are well resolved (each equation is contributing independent information) and that the interfacial fluxes are well resolved, while the reference level velocities are not.

As an example of the quality of these solutions, the reference level velocities and interfacial fluxes corresponding to the relative column weight ($1.8 \times 10^{-4}$) that minimizes the condition number are shown in Figs. 7 and 8. Although the reference level velocities are small and contain little detail, they are still sufficient to correct the geostrophic transport from 211 Sv at section 1 and 170 Sv at section 2 to 188 Sv at both sections, which is much closer to the correct value of 195 Sv. The interfacial fluxes are reproduced remarkably well, both in magnitude and detail (Fig. 8). With this choice of column weights, the inverse problem for interfacial fluxes alone is slightly overdetermined, so that there is enough information to determine fully the interfacial fluxes, plus three pieces of information left over to modify the reference level velocities to get the horizontal fluxes right. The fact that the net horizontal fluxes are well reproduced even with a very smooth representation of the true reference level velocities means that the correlation of velocity and temperature or salinity on small spatial scales does not carry a significant net flux in this region of the FRAM ocean.

In the Tasman Sea box we again reproduce the interfacial fluxes with correlation coefficients greater than 0.9, but the horizontal fluxes have errors of 25%–35%. For example, the geostrophic mass transport is $-7$ Sv at $43\degree S$ and $-9$ Sv at $36\degree S$ relative to a bottom reference level, while the inverse solution mass transport is $-12$ Sv, somewhat closer to the true value of $-18$ Sv. The condition number is a minimum at a relative column weighting of $5.6 \times 10^{-4}$; larger values do not change the solution significantly.

We note that the strategy of choosing the column weight to resolve the interfacial fluxes at the expense of the reference level velocities generally produces the most accurate horizontal fluxes in addition to the most accurate interfacial fluxes: Decreasing the relative weight to redirect information to the determination of
the reference level velocities generally degrades the estimate of the lateral fluxes. For example, a low weight on the interfacial flux unknowns corresponds to the experiment in which the interfacial fluxes were ignored, and the reference level velocities and horizontal fluxes were poorly determined. In the Tasman Sea box, there is a narrow range of larger column weights that give better horizontal fluxes (around 10% errors), but at the expense of having reference level velocities that are as large or larger than, but not correlated with, the true velocities, as well as having poorer interfacial fluxes.

We investigated whether the relatively inaccurate horizontal fluxes in the Tasman Sea might be due to the choice of initial reference level. The depth of minimum kinetic energy (i.e., the sum of squared absolute velocities at each station pair) is about 2800 m in the Tasman Sea region of the model. Choosing the initial reference level at 2800 m (rather than the bottom, which is mostly deeper than 4000 m) led to a slight improvement in the net horizontal fluxes (errors of 21%–33%). Reference levels shallower than 2000 m gave poorer results.

Of the Tasman Sea solutions discussed so far, only the error-free inversion with known interfacial fluxes succeeded in accurately reproducing the horizontal fluxes. Why does the inverse method work less well in the Tasman Sea? With an initial reference level at 2800 m, the model velocity shear gives a transport of $-6.7$ and $-7.4$ Sv across 36° and 43°S, respectively. The true absolute transport across each section is $-17.8$ Sv. [The flow through the Tasman Sea in FRAM is higher than the estimate by Ridgway and Godfrey (1994), based on oceanographic data, of $-8$ Sv, probably because the Indonesian passages are closed in FRAM.] To reproduce the large net transport across the sections, a mean reference level velocity substantially different from zero is required. The least squares solution penalizes departures from the initial guess of zero reference level velocity. In the present inversion, a solution is found that satisfies the constraints but with a net transport of $-13$ Sv. In general, a large “flushing” through a box inverse model will be difficult to reproduce with a least squares solution. These results make the success of the error-free inversion with known interfacial fluxes all the more impressive.

In the Tasman Sea box, the initial state is close to conserving mass overall: The individual layer imbalances are a significant fraction of this total imbalance. Such an imbalance can be rectified by exchanging mass between layers and does not require significant additional net transport through the sections. In the Southern Ocean box, the individual layer mass imbalances are a small fraction of the total mass imbalance so that a reference level velocity with a significant offset from zero is required. The total mass constraint cannot be satisfied by pushing and pulling between interfacial fluxes.
If the total mass transport through the Tasman Sea is known a priori, then this information can be included as a constraint. When this is done, the inverse method does find a solution with the correct mass transport, and the heat and salt anomaly transports are accurate to within 8% of their true values. The interfacial fluxes are little changed from their values without the imposition of the total mass constraint.

The disappointing horizontal flux estimates in the Tasman Sea led us to experiment briefly with a penalty function based on solution smoothness rather than solution size (McIntosh and Veronis 1993). Forcing the solution to be small may not allow the reference level velocities the freedom to give the correct horizontal fluxes. Instead, we penalize a derivative of the reference level velocities while still using a least squares penalty for the interfacial fluxes.

There are four parameters to choose: the order of the derivative, the weight given to the derivative, the weight given to the interfacial fluxes, and the weight given to the equation error. Three of these are independent. We examined parameter space and found that for some settings of these parameters we could obtain a solution that was more accurate than our standard least squares solution. Not only were the horizontal flux errors reduced from 21%–33% to 7%–16%, but the interfacial flux accuracy was improved slightly too.

To obtain this solution, we penalized the first derivative of the reference level velocities, satisfied the equations very accurately, and ensured that the penalty function had roughly equal contributions from the roughness of the reference level velocities and the size of the interfacial fluxes.

The reference level velocities are indeed smooth (in fact, they are nearly uniform across each section) and do not reproduce the spatial structure of the “true” reference level velocity seen in Fig. 3. Penalizing the smoothness rather than the size of \( v \) improves the lateral flux estimates not because the true solution is actually smooth, but because the smoothness constraint permits a solution with a mean value significantly different from zero.

b. Choosing the column weights

We found in the preceding experiments that a relative column weight that minimized the condition number led to good solutions. This empirical finding can be explained by viewing the inverse problem from a statistical point of view. For the SVD solution to be interpreted as a minimum variance solution, the row and column weights must be based on prior estimates of equation and solution standard deviations (Menke 1984; Tarantola 1987). In particular, the relative column weighting will be equal to the ratio of the interfacial velocity standard deviation to the reference level velocity standard deviation. It is possible to calculate from FRAM an estimate of the ratio of standard deviations. For the Southern Ocean box, this value is about \( 1.8 \times 10^{-4} \), which coincides with the value obtained by minimizing the condition number. In the Tasman Sea, the ratio of standard deviations is \( 1.1 \times 10^{-4} \), close to the value obtained by minimizing the condition number (5.6 \( \times 10^{-4} \)). In each case, a relative column weight based on the ratio of standard deviations corresponds to the minimum column weight for which accurate solutions are obtained.

In an inversion of real data, one of course would not know the standard deviation of the true interfacial and reference level velocities. However, an alternative method based on the chi-square statistic provides practical guidance for the choice of column weights. The method is a modification of a method described by Press et al. (1986, p. 503) and is similar to that used by McIntosh and Schahinger (1994). If the elements of the solution have a Gaussian distribution and they are weighted by their standard deviations, then the weighted solution norm has a chi-squared distribution. This norm can be written

\[
\chi^2 = \sum_{i=1}^{N_v} (v_i/\sigma_v)^2 + \sum_{j=1}^{N_w} (w_j/\sigma_w)^2,
\]

where there are \( N_v \) reference level velocities \( v \) with uniform standard deviation \( \sigma_v \), and \( N_w \) effective diapycnal velocities \( w \) with uniform standard deviation \( \sigma_w \). A “good” value of \( \chi^2 \) is equal to the number of degrees of freedom, which for a full rank solution will be something like \( N_v + N_w - M \), where \( M \) is the number of equations. Our assertion is that \( \chi^2 \) is unlikely to attain this value if one of the terms on the right-hand side of (3) is dominant. We choose the ratio of standard deviations, \( \sigma_f/\sigma_r \), so that the ratio of first to second terms on the right-hand side of (3) is \( N_v/N_w \). In other words, the contributions of the reference level velocities and effective diapycnal velocities to \( \chi^2 \) are equal. The relative column weighting is then simply this ratio of standard deviations \( \sigma_f/\sigma_r \). In practice, this involves an iterative procedure in which the relative column weight is chosen, the inverse solution computed, and the ratio of terms calculated. Here we simply scan a range of possible values for the relative column weights, although a minimization technique could be used.

In the Southern Ocean box, the relative column weight obtained from the \( \chi^2 \) procedure is \( 5.3 \times 10^{-4} \), slightly larger than the values obtained from the other two methods (minimum condition number, or based on estimates of standard deviations). In the Tasman Sea box, the \( \chi^2 \) procedure gives a relative column weight of \( 8.7 \times 10^{-4} \), again slightly larger than the values obtained by the other methods. For both boxes, the solutions are very similar for column weights chosen by each of the three methods.

c. Small interfacial fluxes

So far we have studied boxes where the interfacial flux divergence is comparable to the flux divergence
due to the reference level velocity. In the Southern Ocean box, the interfacial flux divergence represents 32% of the total divergence. This figure comes from dividing the rms interfacial mass flux divergence by the sum of this rms and the rms mass flux divergence due to the true reference level velocity. In the Tasman Sea box, the figure is 70%. What happens if the interfacial fluxes are much smaller? This may be the case in a different region of ocean, or maybe FRAM simply overestimates diapycnal fluxes. Do we still obtain good estimates of diapycnal and horizontal fluxes using the same inversion strategies, or does our scheme for selecting the relative column weight “force” the system to introduce large interfacial fluxes?

To test this, we perform inversions where some fraction (denoted by \( \alpha \)) of the true interfacial flux divergence is subtracted from \( b \), thus effectively reducing the size of the interfacial fluxes that the inversion must determine. (For each value of \( \alpha \), the column weighting is chosen according to the \( \chi^2 \) criterion discussed earlier, although we obtain similar results if the column weighting is kept constant at the value that minimizes the condition number of \( A \). If the relative column weighting was chosen according to the standard deviation ratio of the true solution, then it would decrease linearly with increasing \( \alpha \).

The horizontal property fluxes in the Southern Ocean box are insensitive to \( \alpha \), while the interfacial flux correlation coefficients drop below 0.5 when \( \alpha \) is between 0.7 and 0.8, and the fractional rms interfacial flux errors (defined as the rms of the property flux errors divided by the rms of the true fluxes) start to exceed 1.0 in the same interval. (Recall that \( \alpha = 1 \) means the true interfacial fluxes are equal to zero, by construction.) In the Tasman Sea, the horizontal property fluxes are essentially independent of \( \alpha \), having errors of between 20% and 32%, depending on the property. These errors are consistent with those obtained earlier with a relative column weighting chosen to minimize the condition number. The interfacial flux correlation coefficients drop below 0.5 when alpha exceeds 0.6, and the fractional rms interfacial flux errors start to exceed 1.0 at the same time.

As the interfacial flux divergence reduces in importance relative to the barotropic flow divergence, our ability to estimate the interfacial fluxes also decreases. The inversion technique appears to be smart enough to establish that the interfacial fluxes are less important and redirects some information to the reference level velocity. This redirection appears to be only partially dependent on the change in column weighting. When \( \alpha = 1 \), the inverse solution interfacial fluxes should be identically zero. In practice, in the Southern Ocean they have an rms value that is about 16% of the rms of the original interfacial fluxes, while in the Tasman Sea this ratio is 32%. The reference level velocities are substantially unchanged and, in particular, do not gain extra spatial detail. Reducing the column weighting by two orders of magnitude reduces the rms of the interfacial fluxes to about 1% of the original value. Now, the reference level velocities contain almost as much detail as in the case where the interfacial fluxes were assumed known.

We conclude that the \( \chi^2 \) or minimum condition number criteria for selecting a relative column weight produces good solutions for either large or small interfacial fluxes. However, for vanishingly small interfacial fluxes, the column weighting remains too large to recover detail in the reference level velocities.

d. Inversions with errors

We noted previously that the addition of interfacial fluxes to the inversion improved the condition number of the problem considerably (provided the relative weighting of the interfacial fluxes was not too large or too small). The largest and smallest singular values differ by less than two orders of magnitude. Hence, we might expect that including interfacial flux unknowns will make the inversion less sensitive to errors.

Again we add random errors of magnitude 10% of the size of \( b \) and choose the column weighting to minimize the singular value ratio. Despite the presence of errors in \( b \), the inverse solution does not diverge from the true velocity field as the rank is increased (as was seen in the case with no interfacial flux unknowns). In fact, the best solutions in both regions are obtained at full rank, and these solutions have very similar horizontal flux errors to the error-free inversions. The interfacial fluxes look best at full rank, where correlation coefficients greater than 0.9 are found for all properties in both regions (again, much as in the error-free case). If the relative column weighting is chosen according to the \( \chi^2 \) criterion, these results do not change.

To summarize, with the addition of interfacial flux unknowns, the inversion is no longer sensitive to modest errors in \( b \). In particular, the best solutions are obtained at full rank: The interfacial flux unknowns improve the conditioning sufficiently so that there is no longer a need to truncate the number of singular values used.

6. Choosing layers

a. Many layers

In a box inversion where the interfacial fluxes are ignored or considered known, it is possible to choose a number of layers sufficient to over determine the system (provided all equations are independent). If the number of layers is chosen to be greater than the number of station pairs divided by the number of properties conserved, then there will be more equations than unknowns. For the Southern Ocean box, a minimum of 75 layers is required, while for the Tasman Sea, 31 layers are sufficient.

To see if it is possible in practice to over determine
the system in this way, we carried out experiments with 99 layers. This does not necessarily mean we have more resolution than the 32 levels of the FRAM model; because of the general meridional slope of density surfaces, not all surfaces exist at all depths or at both sections.

If the 99-layer inversion is made error free by assuming that the interfacial fluxes and all sources of error are known, then a perfect solution for the reference level velocities is obtained for both boxes! The disadvantage is that the inversion becomes more sensitive to errors in $b$. This sensitivity is due to the small singular values associated with the addition of more layers. Figure 9 shows the singular values for inversions in the Southern Ocean box using 21 and 99 layers. The condition number is increased by more than two orders of magnitude by using 99 layers instead of 21 layers. In the presence of errors, it is more likely to be necessary to reduce the rank of the 99-layer inversion to obtain a robust solution.

With 10% random errors in $b$ and interfacial fluxes known perfectly, increasing the number of layers from 21 to 99 increases the effective rank of the inversion from 33 to 82. (Again we define the effective rank to be the largest rank for which section mass, heat, and salt anomaly transport errors are less than 10%, and the rms reference level velocity is no larger than the rms of the true reference level velocity.) However, the apparent increase in rank in the 99-layer case does not lead to a better solution. The reference layer velocities in the 99-layer case at rank 82 are noisier than for the 21-layer case at rank 33 and do not give a better estimate of the structure of the true reference level velocities.

When the system contains larger errors, as might be expected when interfacial fluxes are ignored, there is no gain to be made by choosing a larger number of layers. The rank necessary to obtain a sensible solution in the Southern Ocean with 21 layers and when interfacial fluxes are ignored was previously established to be below about 8. The singular values of the 21- and 99-layer systems are almost identical out to a rank of about 20, and so we might expect a similarly low rank for the 99-layer case. In practice, a rank of 2 is necessary to obtain a solution with horizontal property flux errors below 10%.

Our conclusion is that when interfacial fluxes are ignored, there is little to be gained by defining more layers than is warranted by the level of error in the equations. If interfacial fluxes are included as unknowns, it is not possible to overdetermine the system. For every layer added, there is an additional interlayer flux for each property to estimate. However, it might be possible to obtain better vertical resolution of the interfacial fluxes.

Figure 10 shows the singular values for inversions in the Southern Ocean box using 21 and 99 layers when interfacial fluxes are included in the inversion as unknowns. Increasing the number of layers has decreased the smallest singular value by an order of magnitude, although at 0.1% of the largest it might still be considered usable if the errors in the equations are not too great.

In the absence of errors, the 99-layer full rank (297 singular values) solution gives very similar results to the comparable 21-layer solution. The only substantial difference is in the improved vertical resolution of the interlayer fluxes.

With 10% random errors in $b$, the same conclusions apply as for the 21-layer inversion. That is, all ranks give accurate horizontal fluxes, but a full rank solution is necessary to obtain accurate interfacial fluxes. The full rank solution is very close to that of the error-free inversion. The order of magnitude increase in the condition number has apparently not made the inversion sensitive to 10% errors.
We conclude that when interfacial fluxes are included in the inversion, the number of layers can be chosen according to the vertical resolution desired.

b. Geopotential layers

Early box inverse models used isopycnals to define layers in which properties were conserved. Because advection and diffusion in the ocean take place mostly along, rather than across, isopycnals, it was believed that the vertical exchange between layers defined by isopycnals would be small and could be ignored. Subsequent inverse modelers have generally continued to use isopycnal or neutral surface layers. However, if interlayer fluxes are included in the box model, it is not necessary to choose surfaces that minimize the interlayer exchange. In this section, we examine the consequences of using geopotential (constant pressure) surfaces rather than neutral surfaces to define our layers. In particular, we choose the FRAM model levels, so that each horizontal layer of computational boxes for tracer points forms one of our box model layers. We use the \( \chi^2 \) criterion to choose the relative column weights. Minimizing the condition number gives similar results.

In the Southern Ocean, the only significant difference when geopotential layers were used was that the interlayer flux correlation coefficient for mass was reduced from 0.96 to 0.74, and the corresponding fractional rms error increased from 0.39 to 0.56. In the Tasman Sea, the horizontal property flux errors increased by about 9 percentage points, while the fractional rms errors for the interlayer fluxes of mass and salt increased from around 0.4 to around 0.7. Surprisingly, the correlation coefficients for all interlayer fluxes stayed at the same high values (about 0.9), and the fractional rms error for the interlayer heat flux decreased from 0.29 to 0.16.

One surprising feature of the dianeutral fluxes in FRAM is that they are the same size or larger than the interlayer fluxes across pressure surfaces. This is true even taking into account Ekman fluxes across outcropping layers. This is clearly an issue that warrants further investigation. For the moment, we mention it merely to indicate that both the isoneutral layer and geopotential layer inversions have significant contributions from the interlayer fluxes.

We conclude that if the inversion contains interlayer fluxes as unknowns, there is no evidence to suggest the solution is improved by using layers defined by neutral surfaces rather than geopotentials. However, the interfacial fluxes may be easier to interpret in the neutral surface framework.

7. Error estimates

We have demonstrated that an appropriately designed box inverse model has some skill in determining lateral and interfacial fluxes. However, we have not addressed the issue of how to assess the accuracy of the inverse method applied to real data, where the velocity field is not known.

The statistical formulation of the inverse problem, sometimes called damped least squares, has the advantage of providing an error estimate for the solution (Menke 1984; Tarantola 1987). The solution is obtained by minimizing

\[
\chi^2 = (x - x_0)^T C_x^{-1}(x - x_0) + e^T C_e^{-1}e, \quad (4)
\]

with respect to the solution \( x \), where \( e = Ax - b \) is the vector of equation errors. The vector \( x_0 \) is strictly the ensemble mean of the solution but is generally taken to be an initial estimate. We have always used the zero vector since we have tried to choose a reference level where the horizontal velocity is small. As far as the interfacial fluxes are concerned, we have no prior information and so set the corresponding elements of \( x_0 \) to zero also. Here \( C_x \) and \( C_e \) are the prior covariance matrices for the solution and equation error respectively. Prior covariance matrices contain statistical information known or estimated before performing the inversion. They are related to the row and column weighting matrices according to

\[
C_i = W_i^2, \quad C_r = W_r^{-2}. \quad (5)
\]

The solution minimizing (4) is

\[
x = x_0 + C_x A^T (A C_x A^T + C_e)^{-1}(b - A x_0), \quad (6)
\]

with posterior covariance \( C_x \), defined by

\[
C_x = C_x - C_x A^T (A C_x A^T + C_e)^{-1}A C_x. \quad (7)
\]

The posterior covariance matrix is an updated version of the prior covariance matrix, which takes into account the introduction of new information represented by the vector \( b \).

We note that the matrix inverse in (6) and (7) is best dealt with by Cholesky decomposition and subsequent solution of two triangular systems of equations in order to avoid inverting a matrix whose condition number is the square of the condition number of \( A \) (Tarantola 1987).

The full rank solution previously obtained using the SVD is reproduced by setting the ratio of interfacial velocity standard deviation to reference level velocity standard deviation to be equal to the relative column weighting, and in the limit as the equation error standard deviation tends to zero.

The statistical interpretation of the inverse problem provides a formal error estimate. We now explore whether this error estimate gives a good practical measure of the accuracy of an inverse solution. The experiment studied is the one in which interlayer fluxes are included as unknowns, and the equations are free of error.

In the Southern Ocean box, we start by guessing the reference level velocity standard deviation to be 0.01 m s\(^{-1}\) and set the effective interfacial velocity standard deviation to 1.8 \( \times 10^{-6} \) m s\(^{-1}\) to give the relative column
FIG. 11. Southern Ocean true and inverse solution reference level velocities together with solution standard deviation. Prior statistics chosen to give previous full rank SVD solution with relative column weighting of $1.8 \times 10^{-4}$.

weighting previously determined to give a good solution by both the condition number and $\chi^2$ criteria. For mass equation error standard deviations less than 0.01 Sv, all solutions are essentially identical, as are the solution covariances. The true and inverse solution together with the solution standard deviation (square root of the diagonal elements of $C_p$) are shown in Figs. 11 and 12. From Fig. 11 it is clear that the reference level standard deviation has not changed from the prior value of 0.01 m s$^{-1}$, indicating that little information has flowed into the reference level velocities. On the other hand, the interfacial flux standard deviations are much smaller than their prior values (typically by a factor of 30) and are also generally smaller than the interfacial flux estimates. The statistical implication is that where our solution has a value in excess of its standard deviation, it is significantly nonzero at the 68% confidence level.

Our estimate of prior solution statistics has so far been quite crude. To see what happens when more accurate prior statistics are available, we use the true solution to estimate a standard deviation for the reference level velocity along each section, and a standard deviation for the effective diapycnal velocity for each property. The only significant change to the results already reported is that the prior (and posterior) standard deviations of the reference level velocities along the section at 147°E are increased from 0.01 to 0.02 m s$^{-1}$. Such a change might have been anticipated from Fig. 11, where the standard deviation for the section at 147°E seems too small.

The consistency between our solution and our estimate of prior statistics may be tested statistically by examining the value of the function we minimize. If we assume that the solution and equation errors have a
Gaussian distribution, then the quantity $\chi^2$ defined by (4) is a chi-squared variable with $\nu = N$ degrees of freedom, where $N$ is the number of equations (Tarantola 1987).

The chi-squared distribution with $\nu$ degrees of freedom has a mean value of $\nu$ and a standard deviation of $\sqrt{2\nu}$. For the Southern Ocean box there are 63 equations, so the chi-squared variable has a mean of 63 and a standard deviation of about 11. Although the chi-squared distribution is skewed, it tends to a Gaussian distribution as the number of degrees of freedom increases. With 63 degrees of freedom there is little difference between the two distributions.

In the Southern Ocean box, with either choice of prior statistics mentioned above, the value of $\chi^2$ is about 38, which is an improbably low value. The interpretation is that we have overestimated our errors slightly (Press et al. 1986; Tarantola 1987). If we reduce all standard deviations uniformly by a factor of 1.3, we obtain a $\chi^2$ value of 63, the most likely value. The implication is that within the uncertainties of estimating errors in the equations and prior solution, our inverse solution is statistically consistent with our prior knowledge.

In the Tasman Sea box, we again estimate the reference level velocity standard deviation to be 0.01 m s$^{-1}$ and set the effective interfacial velocity standard deviation to $5.6 \times 10^{-4}$ m s$^{-1}$ to give the relative column weighting previously determined to give a good solution by the condition number criterion. Similar results are obtained for the Southern Ocean box, although the posterior covariance is a larger fraction of the inverse solution in the Tasman Sea. This is because the prior covariance was larger. The reduction in covariance size was again about a factor of 30.

One unsatisfactory aspect of the Tasman Sea inversion is the $\chi^2$ value of 2.5, which is improbably small compared to 45, the number of degrees of freedom for this inversion. This suggests that the standard deviations are overestimated by a factor of about 4. However, the reference level standard deviation is about right, so the diapycnal velocity standard deviation is the only one that can be reduced. Making this value $1 \times 10^{-4}$ m s$^{-1}$ gives a more reasonable $\chi^2$ value of 28, but degrades the solution by about 10%–20%. The implication is that the statistical inverse method may not give the best solution for a particular inversion, even with accurate prior statistics. However, by design, it will give the solution with the least error over many realizations.

A more careful examination of the dependence of $\chi^2$ on the choice of standard deviations in both inversions indicates just how sensitive this test is. In particular, the assumption that the standard deviation of the diapycnal velocities is independent of depth appears to cause problems. There is considerable depth variation in the diapycnal mass, heat, and salt fluxes in both boxes. Features such as the strong downward flux of heat in the surface layers are likely to be permanent features of the ocean, and prior statistics need to be chosen to reflect this.

Finally, we address the issue of putting error bars on the net horizontal fluxes obtained by the inversion. This is easily done since the fluxes are a linear function of the reference level velocities, and the posterior covariance matrix gives the standard error of the latter. In the Southern Ocean, the standard deviation of the mass flux through the section at 117°E is 12 Sv, while the error in the mass flux at 147°E is 9 Sv. These represent errors of 6% and 4% respectively of the true mass flux. The heat and salt anomaly flux percentage errors were very similar. In the Tasman Sea, mass flux errors were 10 Sv at both sections. In percentage terms, this is a much larger error: 50%–60% of the true mass flux. Similar percentage errors apply for heat and salt anomaly fluxes. These estimates of net flux errors are quite close to the actual errors in the solution for both regions.

We conclude that the statistical interpretation of these types of inverse problems should be viewed with caution if there is insufficient knowledge about the prior statistics. However, if good estimates of prior statistics are available, then inverse solution errors can be useful, both in establishing whether the solution is consistent with the statistics and in estimating horizontal flux errors.

8. Discussion

The experiments described above suggest the following general lessons concerning the performance of the box inverse method.

We find that the FRAM ocean knows about Occam’s Razor. A method that seeks the “simplest” solution by minimizing the size of the correction to an initial model can succeed in reproducing the detailed structure of the reference level velocities in the case where the interfacial fluxes are known, despite the underdetermined nature of the problem.

When inverting real data, the interfacial fluxes will not be known. For accurate solutions, the interfacial fluxes must be included as unknowns in the box inverse model, for two reasons. First, if the interfacial flux divergence is significant in the region considered, a model that does not allow for interfacial fluxes gives poor results. Second, even if the fluxes themselves are small (so that their absence is not a significant source of model error), the near-diagonal nature of the matrix improves the conditioning of the problem, decreasing the sensitivity of the overall problem to data errors.

When interfacial flux unknowns are included, the columns of $\mathbf{A}$ must be properly weighted. We have found empirically that choosing the relative weighting between reference level velocity and interfacial flux unknowns to minimize the condition number leads to the most accurate solutions. By minimizing the condition number, this choice also makes the inversion insensitive to error. The fact that a column weight that minimizes the condition number leads to good solutions can be understood by considering the inverse problem from a
statistical point of view. The optimal choice of relative column weighting is one that ensures that the reference level velocities and interfacial flux unknowns make equal contributions to the chi-squared statistic given in (3). This weight can be found by iteratively choosing a column weight and calculating the ratio of terms in (3) or by applying a minimization technique. In practice, the minimum condition number and chi-squared criteria lead to similar values for the relative column weight.

When column weights are chosen to minimize the condition number, individual interfacial flux unknowns are well resolved, while the individual reference level velocity unknowns are poorly resolved. Nevertheless, the net lateral fluxes are determined accurately in the Southern Ocean box. The successful flux estimates reflect the fact that the variability of the reference level velocity on small spatial scales does not carry a significant net flux of heat or salt. Attempting to force the system to resolve the reference level velocities by decreasing the weight given to the interfacial fluxes degrades both the lateral and interfacial flux estimates. The apparent exception to this rule is if the interfacial fluxes are negligible or known. In this case the detailed structure can be reproduced, but such an inversion is overly sensitive to data error (see section 4).

The lateral fluxes in the Tasman Sea were not determined as accurately as in the Southern Ocean box. The relative lack of success in determining the lateral fluxes in the Tasman Sea box is largely due to a failure to estimate the section-mean reference level velocity, rather than a failure to determine the detailed spatial structure: When the net volume transport across the sections is imposed as a constraint, the heat and salt transports are also well determined, despite the lack of spatial structure in the estimated reference level velocities. In other words, the correlation between changes in barotropic velocity and changes in property concentrations over small spatial scales is sufficiently small that good estimates of lateral fluxes can be determined with a “smoothed” reference level velocity. The improved flux estimates obtained when the roughness rather than the size of $x$ was penalized support this conclusion.

We point out that the lack of success in determining lateral fluxes in the Tasman Sea case is not due solely to the fact that the mean value of the reference velocity is not equal to zero. While it is true that the least squares solution tends to penalize offsets from zero, it does not forbid them: if such a mean offset is required to satisfy the constraints (as in the total mass constraint in the Southern Ocean), then the least squares solution will include one. The poor flux estimates in the Tasman Sea result because there are multiple solutions. Both the SVD and the true solution satisfy the equations exactly, but the true reference level velocity is larger than the simplest solution found by the SVD.

The number and type of surfaces chosen to define the layers can be based on the physical phenomena of interest, provided that interfacial fluxes are included as unknowns. Here we have compared the use of geopotential and isopycral surfaces with from 21 to 99 layers. The reference level velocities are insensitive to the type and number of layers used. The interfacial fluxes are marginally more accurate using isopycnals rather than geopotential surfaces. Note, however, that interfacial fluxes across geopotential surfaces spanning large areas may be difficult to interpret in regions like the Southern Ocean, where the geopotential surfaces may intersect a wide range of density. As the number of layers is changed, the accuracy of the interfacial fluxes is neither improved nor degraded; however, increasing the number of layers increases the “vertical” resolution of the interfacial fluxes.

Our study extends earlier attempts to test inverse models using numerical model output by Bigg (1986) and Killworth and Bigg (1988). The FRAM model differs from the numerical models used by these authors primarily in that it includes both temperature and salinity, includes realistic bottom topography, and is kept close to the observed climatology by the robust diagnostic forcing applied. The inverse methods used are essentially similar. Both Bigg (1986) and Killworth and Bigg (1988) included interfacial fluxes, although neither paper discusses whether the inverse method succeeds in accurately reproducing the interfacial fluxes. Constraints were written only for volume conservation. The question of how to choose the relative column weights when interfacial fluxes are included was not addressed by the earlier authors, nor was the sensitivity of the method to errors in the model or the data. Both Bigg (1986) and Killworth and Bigg (1988) gave the box inverse method a somewhat less encouraging report card than we have presented. Bigg (1986) concluded that inversions using large boxes gave poor results (although this result was apparently retracted in Killworth and Bigg 1988). The later study found that assuming no flow at the bottom usually gave more accurate predictions of the velocity field than any of the inversion techniques they considered. In our experiments, the inverse solution outperformed any choice of initial level of no motion, based on criteria such as the extent to which FRAM’s lateral and interfacial fluxes were reproduced.

Several important issues relevant to inversions of real data have not been addressed by this study. Probably the largest source of error in box inverse models is due to the fact that ocean sections are not synoptic and the ocean circulation is not steady. It is likely that good estimates of the mean ocean circulation will require averaging a number of synoptic snapshots.

The issue of interfacial fluxes requires further attention. In particular, it will be interesting to see if the method can distinguish between interfacial advection and diffusion. The correlation of interfacial advection and property concentration variations about their mean values (the Reynolds term) accounts for a significant fraction of the net interfacial flux in both regions of FRAM we have considered. The extent to which this is
true of the real ocean is unknown. However, the FRAM results suggest that a parameterization of the diffusive flux across interfaces must take the Reynolds flux into account, at least when the interfaces span large areas.

9. Summary

Output from the FRAM numerical model provides a means of testing the ability of a box inverse model to infer reference level velocities and diapycnal fluxes. There are some deficiencies to this form of testing; FRAM may not represent some aspects of the real ocean circulation accurately. However, any inverse model that failed to obtain reasonable results on FRAM output must surely be suspect when applied to real data.

If diapycnal fluxes may be ignored (either because they are insignificant or because an accurate estimate is known), the inverse model does surprisingly well at reproducing the spatial structure of the reference level velocities, despite the underdetermined nature of the problem. However, this inversion is sensitive to error due to its large condition number. In particular, this inversion is sensitive to errors caused by ignoring nonzero diapycnal fluxes. Diapycnal flux divergence in two areas of the FRAM ocean is a significant fraction of the horizontal flux divergence. Ignoring the former leads to worthless inversions.

We include diapycnal fluxes in our inversion by defining a different flux for each property at each layer interface. The inversion is able to estimate these fluxes very accurately for a particular range of column weights. At the same time, the reference level velocities are highly smoothed and contain very little of the detailed structure of the true reference level velocities. There is no choice of column weights that gives accurate spatial detail in the reference level velocities if the diapycnal fluxes are nonzero. In the Southern Ocean box, the horizontal fluxes from the inversion are within 4% of the true values despite the lack of detail in the reference level velocities. The Tasman Sea inversion has larger horizontal flux errors, around 30%, but this is still about half the error before the inversion.

The column weights determine the relative contribution of diapycnal fluxes and reference level velocities to the least squares penalty function. Choosing this relative weighting appropriately is an important factor in the success of an inversion. We find that a good strategy is to choose the relative weighting to minimize the condition number of the matrix $\mathbf{A}$. Relative weights larger than this value tend to give similar solutions. A major advantage of such a strategy is that it makes the inversion as tolerant as possible of data error. The condition number is somewhere between 10 and 100, and the inversions are insensitive to the estimated data errors of around 10%. Hence, we can always use a full rank inversion, eliminating the need to estimate this parameter. We conducted two experiments to determine how important the choice of layers was in obtaining good inverse solutions. Traditional inversions in which vertical exchange of properties between layers is not permitted rely on a clever choice of layers to warrant such an assumption. It might be expected that when interlayer fluxes are allowed that the choice of layers becomes less important. This was indeed the case. We concluded that the number of isopycnal layers can be chosen to obtain the desired vertical resolution. We also found that using geopotential layers led to accurate inversions. One surprising aspect of the latter experiment was that, for all properties and in both regions, the diapycnal fluxes were 2–4 times larger than the vertical fluxes. Without further work, it is impossible to say whether this is indicative of the real ocean or an artifact of the numerical model.

The statistical approach to solving inverse problems was examined and found to provide a satisfactory measure of accuracy for the inverse solution. However, a $\chi^2$ test on the consistency of the solution and the prior statistical assumptions indicated the sensitivity of such an approach to these assumptions.

We find that when the box inverse model is constructed as described above, we can derive accurate estimates of interfacial fluxes in both the Tasman Sea and Southern Ocean. Although one might argue that the interfacial fluxes in FRAM may be larger than in the real ocean, we showed that the method still estimated the interfacial (and lateral) fluxes successfully as the size of the interfacial fluxes was reduced. This is, to our knowledge, the first demonstration of a successful method for calculating net interfacial fluxes over large areas of the ocean. Such fluxes have proven difficult or impossible to measure directly. Accurate estimates of interfacial fluxes are essential to test our ideas about the large-scale dynamics of the ocean circulation. The fact that a properly designed box inverse model can determine interfacial property fluxes is encouraging, and suggests that such methods will prove useful tools for the analysis of the hydrographic sections collected as part of the World Ocean Circulation Experiment.

Acknowledgments. We thank Phil Morgan for writing much of the box model software; David Jackett for providing early versions of the neutral surface labeling software; Susan Wijffels, Nathan Bindoff, and Trevor McDougall for comments on an early draft of the manuscript; and David Webb and Beverly de Cuevas for providing the FRAM data. This work contributes to the DEST Climate Change Research Program.

REFERENCES


Killworth, P. D., and G. R. Bigg, 1988: An intercomparison of inverse...


