J. Fluid Mech. (2005), *vol.* 531, *pp.* 123–157. © 2005 Cambridge University Press
doi:10.1017/S0022112005003812 Printed in the United Kingdom

Generalized semi-geostrophic theory on a sphere

By M. J. P. CULLEN[1], R. J. DOUGLAS[2], I. ROULSTONE[3] AND M. J. SEWELL[4]

[1]Met Office, Fitzroy Road, Exeter EX1 3PB, UK
[2]Department of Mathematics, University of Wales, Aberystwyth SY23 3BZ, UK
[3]Department of Mathematics and Statistics, University of Surrey, Guildford GU2 7XH, UK
[4]Department of Mathematics, University of Reading, Reading RG6 6AX, UK

(Received 3 February 2003 and in revised form 20 December 2004)

It is shown that the solution of the semi-geostrophic equations for shallow-water flow can be found and analysed in spherical geometry by methods similar to those used in the existing f-plane solutions. Stable states in geostrophic balance are identified as energy minimizers and a procedure for finding the minimizers is constructed, which is a form of potential vorticity inversion. This defines a generalization of the geostrophic coordinate transformation used in the f-plane theory. The procedure is demonstrated in computations.

The evolution equations take a simple form in the transformed coordinates, though, as expected from previous work in the literature, they cannot be expressed exactly as geostrophic motion. The associated potential vorticity does not obey a Lagrangian conservation law, but it does obey a flux conservation law, with an associated circulation theorem.

The divergence of the flow in the transformed coordinates is primarily that naturally associated with geostrophic motion, with additional terms coming from the curvature of the sphere and extra ‘curvature’ resulting from the variable Coriolis parameter in the generalized coordinate transformation. These terms are estimated, and are found to be very small for normal data. The estimate is verified in computations, confirming the accuracy of the local f-plane approximation usually made with semi-geostrophic theory.

1. Introduction

Quasi-geostrophic theory has for a long time been the most widely used model of large-scale atmospheric circulations. This is because of its conceptual simplicity, and the possibility of finding analytic solutions. However, the quasi-geostrophic approximation uses a constant Coriolis parameter in the definition of the geostrophic wind, and a fixed reference state static stability that is independent of horizontal position. Neither of these approximations is valid on large scales in the atmosphere, though that does not prevent the solutions from being conceptually useful. The Type II geostrophic approximation identified by Phillips (1963) does not use a reference state static stability or constant Coriolis parameter, and is thus valid on large scales. However, inertial effects are neglected entirely, which is too severe an approximation for most purposes in the atmosphere. The geostrophic momentum approximation, originally introduced by Eliassen (1948), and developed and promoted by Hoskins (1975), allows the use of the correct variation of the Coriolis parameter and the static stability. This allows the advantages of both types of geostrophic approximation to be subsumed into a single set of equations. In order to achieve this, while retaining energetic consistency, the geostrophic approximation is only made in the calculation of the momentum, not in the fluid trajectory. This particular feature of the approximation is now well understood in terms of Hamiltonian mechanics. The resulting semi-geostrophic equations can thus describe a number of subsynoptic flows such as fronts embedded in cyclones, and interactions of large-scale flow with topography and convection. These were reviewed by Cullen *et al.* (1987). However, explicit solutions of the equations have normally been obtained only with a constant Coriolis parameter, where the geostrophic coordinate transformation of Hoskins reduces the equations to a similar form to the quasi-geostrophic equations. The geometric solution procedure introduced by Cullen & Purser (1984) also relies on a constant Coriolis parameter. The solutions have proved conceptually useful despite this restriction.

The semi-geostrophic equations have, however, been integrated numerically on the sphere, without using a coordinate transformation. Mawson & Cullen (1992) showed that ageostrophic cross-equatorial flows can be predicted as a response to suitably imposed forcing. Mawson (1996) showed, using a shallow-water version of the equations, that the model supports the same Rossby wave solutions as the full equations, as long as the geostrophic wind satisfies the inertial stability condition. The inertial stability condition severely constrains the permitted solutions close to the equator. The semi-geostrophic approximation thus contains the 'weak temperature gradient' approximation which has recently become popular in tropical studies, e.g. Polvani & Sobel (2002). Schubert *et al.* (1991) showed, using a zonally symmetric form of the model, that many aspects of the observed Hadley circulation can be simulated. These studies confirm the appropriateness of the semi-geostrophic model for large-scale flows. Cullen (2000) provided theoretical support for this by verifying that the errors in large-scale semi-geostrophic solutions on the sphere decrease as the square of (L_R/L), where L is the horizontal length scale and L_R the Rossby deformation radius. This does not contradict the well-known inaccuracy of the semi-geostrophic treatment of vorticity dynamics, because in the regime where the Rossby number is small and $L_R < L$, vorticity dynamics is less important.

In order to confirm the suitability of the semi-geostrophic model as a simplified model on the sphere, it is necessary to show that the solutions make sense so that, in particular, the geostrophic winds and geostrophic energy remain bounded as the equator is approached. This requires the horizontal pressure gradients to tend to zero at the equator. While this was always the case in the studies referred to above, it is necessary to demonstrate that the required behaviour emerges naturally from the solution procedure. One of the main achievements of this paper is to do this.

A major advantage of the f-plane semi-geostrophic equations is the existence of a robust solution procedure based on the transport of a single scalar, the potential vorticity, followed by inversion of the potential vorticity to obtain the remainder of the variables. Hoskins, McIntyre & Robertson (1985) showed that this is a generic procedure applicable to a number of balanced models. In the semi-geostrophic case, the procedure is facilitated by a coordinate transformation, often referred to as the geostrophic momentum transformation, and the potential vorticity is always invertible. Benamou & Brenier (1998) and Cullen & Gangbo (2001) have exploited this fact to prove that the f-plane semi-geostrophic equations can be integrated for large times from suitable initial data. As discussed by McIntyre & Roulstone (2002), most other balanced models also have solvability conditions which are not always satisfied.

In order to exploit the validity of semi-geostrophic theory on large scales, it is necessary to extend the robust solution procedure from the f-plane to spherical geometry. This is another achievement of this paper. A number of previous attempts have been made; however, all of these have required altering the equations in some way. Salmon (1985) defined a set of equations, directly in a space obtained from a simple transformation of the physical variables, which are Hamiltonian and conserve potential vorticity. He showed that the equivalent equations in the physical space are not the same as the semi-geostrophic equations, but that the change to the equations is within the error made by using the semi-geostrophic equations as an approximation to the primitive equations. Magnusdottir & Schubert (1991), and Purser (1999), both approximated the semi-geostrophic equations in a way that assumes that the flow is approximately zonal, and then showed that the resulting equations can be solved by a coordinate transformation. Shutts (1989) constructed a Hamiltonian semi-geostrophic system for the sphere by regarding the spherical shell as a subset of general three-dimensional space. This leads to the 'planetary' semi-geostrophic system, which recognizes that the axis of rotation of the Earth is the special direction, rather than the local vertical. However, this model does not reduce to the local f-plane model on small regions of the Earth's surface. In view of the success of local f-plane models, and the belief that the spherical geometry will not fundamentally alter their results, we seek a version of the solution procedure that can be applied to the unmodified semi-geostrophic equations in spherical geometry. The approach adopted in this paper is based on preserving the form of the equations of motion in physical Lagrangian variables – with a variable Coriolis parameter – while generalizing the coordinate transformation in such a way that the f-plane geostrophic momentum transformation of both Eliassen (1948) and Hoskins (1975) is recoverable in the limit of a constant Coriolis parameter.

Cullen & Purser (1989) showed that the potential vorticity inversion procedure for f-plane semi-geostrophic theory could be interpreted as a minimization of the energy under the constraint of given inverse potential vorticity, where the inverse potential vorticity is defined as the Jacobian of the mapping from geostrophic and isentropic coordinates to physical coordinates. Hereinafter we use the term 'potential density' for the inverse potential vorticity. In this paper, we generalize Cullen & Purser's (1989) result to the shallow-water case on the sphere. The method used was first introduced by Cullen & Douglas (1998). The first step is to find a coordinate transformation on the sphere that generalizes the geostrophic coordinate transformation, and allows the potential density to be defined. A theorem by McCann (2001) can then be used to show that the potential density can be uniquely inverted, subject to a regularization of the problem at the equator. We demonstrate the 'potential density inversion' in a computation. We show in §4 that the resulting depth field satisfies a concavity condition which is equivalent to a local inertial stability condition. It therefore ensures that the depth field is flat enough near the equator for the solutions to make sense. This condition is the same as the ellipticity condition required in the solution procedure of Mawson (1996). It is also the analogue of the convexity principle used by Cullen & Purser (1984) for the f-plane case.

We then proceed to show that this method allows the robust solution of the semi-geostrophic equations on the sphere. We derive the time evolution equation in the new coordinates. The potential density is transported by a velocity in the transformed space which is in the same direction as the geostrophic velocity, but with magnitude modified by terms that result both from the curvature of the sphere itself and from the variable Coriolis parameter. The local mass conservation equation in physical space transforms to a circulation theorem, so that the integral of the potential density within any material circuit in transformed space is conserved. The divergence of the velocity in transformed space is dominated by the variations of the Coriolis parameter. As material circuits move towards the equator, they expand and the potential density decreases.

Given an initial potential density, the equations can be discretized in time by first solving the energy minimization problem and calculating the depth field and geostrophic winds. These are guaranteed to be inertially stable. An elliptic problem is solved in physical space to calculate the velocity to be used in the transformed space. The transport equation in transformed space is then integrated for a time step. By regularizing the transformation at the equator, we can use the concavity condition on the depth field to show that the procedure converges as the time step is refined. The limit solution will be the solution of a regularized problem. We show in § 2 that the inertial stability condition constrains the depth field to be very flat near the equator. As a result, we can show that a well-defined solution to the original problem is obtained as the regularization is removed.

Finally, we illustrate the time-dependent solutions. In particular, we show that the local f-plane approximation to the potential vorticity is almost conserved, to the extent that it is not clear whether the non-conservation is analytic or numerical. This is because the divergence of the velocity in transformed space can be almost exactly removed by a rescaling of the transformed sphere. The remaining terms are shown to be small for realistic velocities, such as those used in the computations. Thus, a diagnostic based on a potential vorticity calculation in real space will still be useful.

The structure of the rest of the paper is as follows. In § 2, we set out the shallow-water semi-geostrophic equations on the sphere. The f-plane theory works by showing that the equations can be interpreted as describing an evolution through a sequence of minimum energy states. We demonstrate that this interpretation also holds for the spherical case. We derive the conditions for the energy to be minimized, rather than just made stationary, and show that this requires the positive definiteness of a particular matrix. The condition has a similar form to that in the f-plane theory, and can be interpreted as a local inertial stability condition. We write down a formal solution procedure in physical space, and show that there is a solvability condition which is identical to the inertial stability condition. In § 3, we show that the energy minimization problem in the f-plane case can be formulated in terms of potential density inversion. The inversion procedure defines a mapping between geostrophic and physical coordinates which minimizes a rescaled Euclidean distance between the coordinates, subject to given potential density. We generalize this to the sphere by replacing the rescaled Euclidean distance by a distance function defined by rescaling the metric on the sphere with the Coriolis parameter. We analyse this using standard techniques from the calculus of variations, and show that it leads to a generalization of the f-plane geostrophic coordinate transformation. In § 4, we show that the potential density inversion problem can be uniquely solved on the sphere. We show that the solution of this problem is also a solution of the energy minimization problem, and that any solution of the energy minimization problem is the solution of a potential density inversion problem, for some potential density. We exploit this to show that the semi-geostrophic equations can be solved for finite times on the sphere, pointing out some technical issues which require resolution before a rigorous proof can be made. In § 5, we show that the semi-geostrophic equations can be written as a transport equation for the potential density in geostrophic coordinates, with a transport velocity parallel to the geostrophic wind. However, the transport velocity cannot be determined without complete knowledge of the solution, unlike the f-plane case. We show that, if the geostrophic coordinate space is further rescaled, the transport velocity is 'almost' non-divergent. Thus, there is an exact circulation theorem for potential density in geostrophic coordinates, but Lagrangian conservation is only approximate. We demonstrate the coordinate transformation and the solutions in computations.

2. Basic theory

2.1. *The semigeostrophic equations*

The kinematics and dynamics of shallow-water theory on a plane, and its semi-geostrophic approximation, are discussed by Roulstone & Sewell (1996, 1997), for example, and we adopt their notation in quoting some of the equations. Let $\boldsymbol{r}$ denote the position vector of a generic point, from a fixed origin in three-dimensional Euclidean space, on the surface of the sphere of radius a. An increment along the surface can be written in physical components as $\mathrm{d}\boldsymbol{r} = a\cos\phi\,\mathrm{d}\lambda\boldsymbol{i}_\lambda + a\,\mathrm{d}\phi\boldsymbol{i}_\phi = (a\cos\phi\,\mathrm{d}\lambda, a\,\mathrm{d}\phi)$, with orthogonal unit vectors $\boldsymbol{i}_\lambda$ and $\boldsymbol{i}_\phi$ parallel to the coordinate circles of increasing longitude λ and latitude ϕ, respectively.

The motion of a typical particle in shallow-water theory on a sphere can be described by expressing the current Eulerian coordinates of the particle on the surface of the sphere

$$\lambda = \lambda(\alpha, \beta, t), \quad \phi = \phi(\alpha, \beta, t), \tag{1}$$

as functions, on the right-hand sides, of the particle labels (or Lagrangian coordinates) α, β and the time t, such that $\alpha = \lambda(\alpha, \beta, 0)$, $\beta = \phi(\alpha, \beta, 0)$. Incompressibility requires that

$$\frac{h(\alpha, \beta, 0)}{h(\alpha, \beta, t)} = \frac{\partial(\lambda, \phi)}{\partial(\alpha, \beta)}\frac{\cos\phi}{\cos\beta}, \tag{2}$$

where h is the fluid depth at the particle position. We denote the right-hand side of (2) by j and assume $0 < j < \infty$. This makes available the inverse description $\alpha = \alpha(\lambda, \phi, t)$, $\beta = \beta(\lambda, \phi, t)$ of the motion (1), and allows us to transfer between Lagrangian and Eulerian descriptions whenever required. In particular, we can express $h(\alpha, \beta, t)$ as a different function $h(\lambda, \phi, t)$. Within the fluid, we will assume $h > 0$, though noting that semi-geostrophic theory can also describe situations where $h = 0$ over part of the domain (see Cullen & Purser 1989). Unless otherwise stated, we shall use the same letter to denote a function and its generic values, as just illustrated.

If the coordinates are in a frame of reference rotating with the Earth, the particle acceleration $(a(\cos\phi\ddot{\lambda} - 2\sin\phi\dot{\phi}\dot{\lambda}), a(\ddot{\phi} + (\dot{\lambda})^2\sin\phi\cos\phi))$ has an additional term $(-fa\dot{\phi}, fa\cos\phi\dot{\lambda})$, where the Coriolis parameter $f = 2\Omega\sin\phi$ is the component of the angular velocity vector normal to the surface and Ω is the spin of the Earth. The superposed dots signify the Lagrangian time derivatives, following the particle. (Some authors write $\mathrm{D}^n/\mathrm{D}t^n$ for these derivatives.) The continuity equation can be written as the time derivative of (2) following the particle,

$$\dot{h} + h\nabla\boldsymbol{\cdot}\dot{\boldsymbol{r}} = 0. \tag{3}$$

where $\dot{\boldsymbol{r}} = (a\cos\phi\dot{\lambda}, a\dot{\phi})$ is the particle velocity.

The shallow-water momentum equations on the sphere can then be written, using the spherical polar Eulerian coordinates defined above, as

$$a(\cos\phi\ddot{\lambda} - 2\sin\phi\dot{\phi}\dot{\lambda}) - fa\dot{\phi} + \frac{g}{a\cos\phi}\frac{\partial h}{\partial\lambda} = 0, \tag{4a}$$

$$a(\ddot{\phi} + (\dot{\lambda})^2\sin\phi\cos\phi) + fa\cos\phi\dot{\lambda} + \frac{g}{a}\frac{\partial h}{\partial\phi} = 0. \tag{4b}$$

The constant acceleration due to gravity is denoted by g, and the depth at a fluid particle position has now been written as a function $h(\lambda, \phi, t)$. Define the geostrophic wind (u_g, v_g) to have local physical components

$$u_g = -\frac{g}{fa}\frac{\partial h}{\partial \phi}, \tag{5a}$$

$$v_g = \frac{g}{fa\cos\phi}\frac{\partial h}{\partial \lambda}. \tag{5b}$$

The geostrophic momentum approximation to (4a) and (4b) is then

$$\dot{u}_g - \dot{\lambda} v_g \sin\phi - fa\dot{\phi} + \frac{g}{a\cos\phi}\frac{\partial h}{\partial \lambda} = 0, \tag{6a}$$

$$\dot{v}_g + \dot{\lambda} u_g \sin\phi + fa\cos\phi\dot{\lambda} + \frac{g}{a}\frac{\partial h}{\partial \phi} = 0. \tag{6b}$$

Equations (3), (5) and (6) are the semi-geostrophic system to be solved, but at present there are no results establishing existence and uniqueness properties for these equations because the Coriolis parameter is a function of position. It is also not clear whether the equations make sense as the equator is approached. We seek to establish a solution procedure both within a given domain D of the surface of the sphere and on the sphere as a whole. In the first case, it is a basic assumption that particles cannot enter or leave D across the boundary.

2.2. *Conservation of energy and potential vorticity*

In f-plane semi-geostrophic theories, it is easy to show that the total energy, which is the sum of a geostrophic kinetic energy and a potential energy, is conserved following the motion of the fluid particles. Furthermore, also in the case of f-plane theories, the potential vorticity is a Lagrangian conserved quantity. These issues are discussed in some detail in Roulstone & Sewell (1997) and in McIntyre & Roulstone (2002); the latter paper also explains in some detail why the geostrophic flow, and not the actual particle motion, appears in the conserved quantities.

When the Coriolis parameter becomes a function of position as it is in (6), no form of potential vorticity conservation is known to exist (e.g. see the discussion of shallow-water semi-geostrophic theory in Roulstone & Sewell (1996), § 3), save by making the approximations discussed by Salmon (1985), Shutts (1989) and Magnusdottir & Schubert (1991). These approximations amount to altering the original equations (3), (5) and (6). However, as we shall now demonstrate, we can establish a conservation law for the total energy on the assumption that the energy remains finite as we approach the equator (i.e. when f vanishes).

The total geostrophic kinetic plus potential energy, associated with the geostrophic wind, is defined by

$$G = \int \left(\tfrac{1}{2}h\left(u_g^2 + v_g^2\right) + \tfrac{1}{2}gh^2\right) \mathrm{d}\Sigma, \tag{7}$$

where $\mathrm{d}\Sigma = a^2\cos\phi\,\mathrm{d}\lambda\,\mathrm{d}\phi$ is the area element of the sphere, and the integration is either over a simply connected domain D of the sphere, or over the whole sphere. In discussing boundary conditions, we now assume that D is a finite closed region of the sphere, possessing a boundary. Then G has the property that

$$\dot{G} = -\tfrac{1}{2}\int g\nabla\cdot(h^2\dot{\boldsymbol{r}})\,\mathrm{d}\Sigma = -\tfrac{1}{2}\oint gh^2\boldsymbol{n}\cdot\dot{\boldsymbol{r}}\,\mathrm{d}s, \tag{8}$$

where ds is the line element along the boundary of D. The first equality in (8) is a consequence of (6) and continuity (one form of which is that $h\,\mathrm{d}\Sigma$ is constant), before any boundary conditions are used. Here $\boldsymbol{n}$ denotes the outward unit normal to the boundary of D. Thus $\boldsymbol{n}$ is tangential to the sphere. The second equality in (8) depends on a 'divergence theorem' on the sphere. The sphere over which the integration is performed is considered as being embedded in a three-dimensional space, so that the form of the divergence theorem we require is subtly different from that usually found in standard textbooks (although it is easily formulated in the language of tensor calculus) and we therefore furnish a proof of this result, for completeness, in the Appendix.

The foregoing equations, and (8) in particular, imply the following result.

Theorem 1. $\dot{G}=0$ *when* (6) *with continuity holds within D, together either with* $\boldsymbol{n}\cdot\dot{\boldsymbol{r}}=0$ *on the boundary of D, or with integration carried out over the whole sphere so that* $\int\nabla\cdot(h^2\dot{\boldsymbol{r}})\,\mathrm{d}\Sigma=0$. *Thus G is conserved as an overall property of the semi-geostrophic flow, even in spherical geometry.*

The remainder of this paper is devoted to the presentation and discussion of a method for establishing the existence of a set of solutions of the semi-geostrophic equations (3), (5) and (6) on the sphere for which the conservation of energy and the evolution of potential vorticity are well defined. In particular, we show that there is a set of solutions for which the energy integral (7) is bounded, implying that the geostrophic winds are bounded at the equator.

2.3. *Identification of geostrophic balance with a stationary energy state*

We now show that solutions of the semi-geostrophic equations in spherical geometry are characterized by being minimum energy states, as in the f-plane case, in a sense to be made precise as follows. With any shallow motion of local depth h on the sphere, we can associate a vector field having physical components (u, v) (say). Thus the vector is $u\boldsymbol{i}_\lambda + v\boldsymbol{i}_\phi$. It can be thought of as a velocity, but it need not have that interpretation, which is therefore purely notional, to suggest possible physical consequences. By analogy with (7) we can then define a notional energy

$$E=\int\left(\tfrac{1}{2}(u^2+v^2)+\tfrac{1}{2}gh\right)h\,\mathrm{d}\Sigma. \tag{9}$$

This is a functional of u, v and h, regarded as functions of position over Σ, which has the following property.

Theorem 2. *The conditions for the integral E to be stationary with respect to variations satisfying continuity* $\delta(h\,\mathrm{d}\Sigma)=0$ via

$$\delta h=-h\nabla\cdot\delta\boldsymbol{r} \tag{10}$$

in D, where D is (part of) the sphere, and

$$\delta u=fa\delta\phi+v\sin\phi\delta\lambda,\quad \delta v=-fa\cos\phi\delta\lambda-u\sin\phi\delta\lambda \tag{11}$$

together with

$$h\boldsymbol{n}\cdot\delta\boldsymbol{r}=0 \tag{12}$$

on the boundary of D as necessary, are that

$$u=u_g,\quad v=v_g. \tag{13}$$

The stationary value of E is G.

Proof. The calculation is formally similar to that which delivers (8) above. Using (10) first, followed by the divergence theorem for (part of) the sphere, we obtain

$$\delta E = \int (u\delta u + v\delta v + g\delta \boldsymbol{r}\cdot\nabla h)h\,\mathrm{d}\Sigma - \tfrac{1}{2}g\oint h^2\boldsymbol{n}\cdot\delta\boldsymbol{r}\,\mathrm{d}s. \tag{14}$$

Using (11) and (5), with (12) when required, we obtain

$$\delta E = \int (fa\delta\phi(u-u_g) - fa\cos\phi\delta\lambda(v-v_g))h\,\mathrm{d}\Sigma. \tag{15}$$

Then, for E to be stationary with respect to arbitrary variations $\delta\phi, \delta\lambda$, we must require (13) to hold. □

The substance of the result is that E is stationary when the notional velocity $\boldsymbol{u} = (u, v)$ is equal to the geostrophic wind within the fluid, and when the boundary conditions are satisfied. The choice of variations in (11) represents the effect of a notional displacement in a rotational system where the effect of any pressure perturbation generated by the displacement is neglected, as comparison of (11) with (6) shows. The increments in (11) are definitions, and in (10) δh is deduced from $\delta\boldsymbol{r}$ using incompressibility.

Shutts & Cullen (1987) analyse the physical significance of E being minimized, rather than just stationary, for the case of constant f. They show that it corresponds to the stability of a geostrophic state, viewed as a solution of the full primitive equations, to perturbations of the form

$$\delta u = f\delta y, \quad \delta v = -f\delta x, \quad \nabla\cdot(\delta x, \delta y) = 0, \tag{16}$$

which are the analogues of (10) and (11) in plane geometry. They also discuss the validity of the assumption that pressure perturbations can be neglected. They show (pp. 1321–1323) that it is valid if the basic flow and perturbations both satisfy the assumptions of semi-geostrophic theory, i.e. that one horizontal length scale is large.

We therefore derive necessary conditions for E to be minimized under the variations (11), closely following the method of Shutts & Cullen (1987). Rewrite (15), using $\delta\boldsymbol{r} = (a\cos\phi\delta\lambda, a\delta\phi)$, as

$$\delta E = \int \delta\boldsymbol{r}\cdot(-f(v-v_g), f(u-u_g))h\,\mathrm{d}\Sigma. \tag{17}$$

Then, taking a second variation,

$$\delta^2 E = \int \delta(f\delta\boldsymbol{r}\cdot(-(v-v_g), u-u_g))h\,\mathrm{d}\Sigma, \tag{18}$$

and since $\boldsymbol{u} = \boldsymbol{u}_g$ when $\delta E = 0$, this reduces to

$$\delta^2 E = \int f\delta\boldsymbol{r}\cdot(-\delta(v-v_g), \delta(u-u_g))h\,\mathrm{d}\Sigma. \tag{19}$$

Substituting for $\delta\boldsymbol{u}$ from (11) and using $\boldsymbol{u} = \boldsymbol{u}_g$ gives

$$\delta^2 E = \int f\delta\boldsymbol{r}\cdot((fa\cos\phi\delta\lambda + u_g\sin\phi\delta\lambda, fa\delta\phi + v_g\sin\phi\delta\lambda) + \delta(v_g, -u_g))h\,\mathrm{d}\Sigma. \tag{20}$$

Equation (5) gives $(v_g, -u_g) = gf^{-1}\nabla h$. Thus $\delta(v_g, -u_g) = g\delta(f^{-1}\nabla h)$. If we write ∂ for a change at a fixed position in space caused by a displacement, then

$$g\delta(f^{-1}\nabla h) = g\partial(f^{-1}\nabla h) + g\delta\boldsymbol{r}\cdot\nabla(f^{-1}\nabla h) = g\partial(f^{-1}\nabla h) + \delta\boldsymbol{r}\cdot\nabla(v_g, -u_g). \tag{21}$$

Since $\partial f = 0$ and (10) implies that $\partial h = -\nabla \cdot (h\delta \boldsymbol{r})$, we have

$$g\delta(f^{-1}\nabla h) = -gf^{-1}\nabla(\nabla \cdot (h\delta \boldsymbol{r})) + \delta \boldsymbol{r} \cdot \nabla(v_g, -u_g). \tag{22}$$

Substituting (22) into (20) and integrating by parts gives

$$\delta^2 E = \int f\delta \boldsymbol{r} \cdot ((fa\cos\phi\delta\lambda + u\sin\phi\delta\lambda,\ fa\delta\phi + v\sin\phi\delta\lambda) \\ + a(\cos\phi\delta\lambda, \delta\phi) \cdot \nabla(v_g, -u_g))h + g\,(\nabla \cdot (h\delta \boldsymbol{r}))^2\ \mathrm{d}\Sigma. \tag{23}$$

The second term is positive definite. The condition for the energy to be minimized is therefore that the first term is positive definite. Writing it in the form $\delta \boldsymbol{r} \cdot \boldsymbol{P} \cdot \delta \boldsymbol{r}$, the condition is that the matrix

$$\boldsymbol{P} = f \begin{pmatrix} f + \dfrac{1}{a\cos\phi}\dfrac{\partial v_g}{\partial\lambda} + \dfrac{u_g\tan\phi}{a} & \dfrac{\partial v_g}{\partial\phi} \\ -\dfrac{1}{a\cos\phi}\dfrac{\partial u_g}{\partial\lambda} + \dfrac{v_g\tan\phi}{a} & f - \dfrac{1}{a}\dfrac{\partial u_g}{\partial\phi} \end{pmatrix} \tag{24}$$

is positive definite.

We can see that this is the standard semi-geostrophic form of the inertial stability condition found by Shutts & Cullen (1987) using the local value of f and written in spherical polar coordinates. The derivatives of f do not enter the condition. To see what effect this condition has at the equator, we calculate the terms on the diagonal of $\boldsymbol{P}$,

$$\left.\begin{aligned} &\frac{1}{a\cos\phi}\frac{\partial v_g}{\partial\lambda} = \frac{g}{2\Omega a^2\sin\phi\cos^2\phi}\frac{\partial^2 h}{\partial\lambda^2}, \\ &\frac{u_g\tan\phi}{a} \simeq 0, \\ &-\frac{1}{a}\frac{\partial u_g}{\partial\phi} = \frac{g}{2\Omega a^2}\left(-\frac{\cos\phi}{\sin^2\phi}\frac{\partial h}{\partial\phi} + \frac{1}{\sin\phi}\frac{\partial^2 h}{\partial\phi^2}\right). \end{aligned}\right\} \tag{25}$$

Thus the condition that $f(f - (1/a)(\partial u_g/\partial\phi)) > 0$ requires that

$$g\frac{\partial h}{\partial\phi} \simeq 2\Omega U_0\sin\phi + O(\phi^3) \tag{26}$$

for some constant U_0. Thus $u_g = U_0 + O(\phi^2)$. The condition that $f(f+(1/a\cos\phi)(\partial v_g/\partial\lambda)) > 0$ means that $\partial h/\partial\lambda = O(\phi^2)$. The implied conditions on h are much more severe than those required for (u_g, v_g) to be finite at the equator.

2.4. *Solution procedure in physical space*

We now describe the method used to solve the semi-geostrophic equations on the sphere in physical space. The numerical model of Mawson (1996) is based on this procedure. Assume we are given initial data h defined over the whole sphere or over some bounded subset D of it. Calculate the geostrophic wind from h using (5), and call the result $(u_g(0), v_g(0))$. Assume that these data are such that the associated geostrophic energy $G(0)$ calculated using (7) is finite, and calculate $\boldsymbol{P}$ using $\boldsymbol{u}_g(0)$ in (24). We can then write (3) and (6), following Schubert (1985), as:

$$\left.\begin{aligned} &\boldsymbol{P}\dot{\boldsymbol{r}} + \frac{\partial}{\partial t}\nabla h = -f^2\boldsymbol{u}_g, \\ &\frac{\partial h}{\partial t} + \nabla \cdot (h\dot{\boldsymbol{r}}) = 0. \end{aligned}\right\} \tag{27}$$

These equations can be combined to give

$$\frac{\partial h}{\partial t} - \nabla \cdot \left(h\boldsymbol{P}^{-1} \frac{\partial}{\partial t} \nabla h \right) = \nabla \cdot (h\boldsymbol{P}^{-1} f^2 \boldsymbol{u}_g). \tag{28}$$

Equation (28) is a Helmholz equation for $\partial h/\partial t$ provided that the matrix $\boldsymbol{P}$ is sign definite. If $\boldsymbol{P}$ is positive definite, the eigenvalues of the principal part of the Helmholz operator will all be positive, and we can expect (28) to have a unique solution for $\partial h/\partial t$. Equation (27) can then be integrated forwards in time.

In the f-plane case, Cullen & Gangbo (2001) prove that the energy minimization condition derived formally in the preceding subsection is exactly the condition required for (27) to be solvable (in a suitable sense). The method used is to show that the equations can be rewritten in transformed (geostrophic) coordinates as a transport equation for $(\det \boldsymbol{P})^{-1}$, the potential density. At each time, the depth field and geostrophic winds are calculated from the potential density by solving an energy minimization problem equivalent to that described in the previous subsection. It is proved that this requires the matrix $\boldsymbol{P}$ to be positive definite. In § 4, we demonstrate that the arguments of Cullen & Gangbo can be applied to the spherical case.

3. The geostrophic momentum transformation on the sphere

3.1. *Solution of the energy minimization problem using geostrophic coordinates in the f-plane case*

In the f-plane case, Cullen & Purser (1989) showed that the problem of minimizing E subject to the variations (16) could be solved uniquely. They gave an intuitive proof of this, which has since been made rigorous by Douglas (1998) for incompressible, three-dimensional stratified semi-geostrophic flows with rigid boundaries, and by Cullen & Gangbo (2001) for shallow-water semi-geostrophic flow in a bounded region. In both cases, the boundary conditions were that no fluid enters or leaves the region across the boundaries. We describe the procedure in a region D using Cartesian coordinates (x, y). A key step is to rewrite the notional velocity as

$$(u, v) = f(y - Y', X' - x) \tag{29}$$

in terms of a new pair of generic Cartesian coordinates X', Y'. The class of variations (16) and (10) under which the energy is to be minimized now take the form

$$\left.\begin{aligned} \delta X' = \delta Y' = 0, \\ \delta h = -h\nabla \cdot \delta \boldsymbol{r}. \end{aligned}\right\} \tag{30}$$

These together imply that

$$\delta\sigma = 0, \tag{31}$$

where

$$\sigma = h\frac{\partial(x, y)}{\partial(X', Y')}. \tag{32}$$

Equation (32) states that σ is the mass density in (X', Y') space. It is thus a non-negative quantity as long as the transformation between (x, y) coordinates and (X', Y') coordinates remains well-defined. A state of rest with h equal to a uniform value h_0 corresponds to $\sigma = h_0$. Any other choice of σ implies some excess energy above the rest state.

Using (29), the notional kinetic energy term in (9) can be written as

$$\tfrac{1}{2}f^2 \int ((X'-x)^2 + (Y'-y)^2) h \, \mathrm{d}x \, \mathrm{d}y, \tag{33}$$

which may thus be regarded as a weighted integrated distance between the physical position $\boldsymbol{x}$ and another associated point $\boldsymbol{X}' = (X', Y')$ in the Euclidean space. It is therefore minimized by making the $\boldsymbol{X}'$ points correspond as closely as possible to the physical positions. This is the key standpoint from which our generalizations will follow.

If we define a distance $d(\boldsymbol{x}, \boldsymbol{X}')$ between $\boldsymbol{x}$ and $\boldsymbol{X}'$ in the f-plane to be such that its square is

$$d(\boldsymbol{x}, \boldsymbol{X}')^2 = f^2((X'-x)^2 + (Y'-y)^2), \tag{34}$$

the notional energy (9) can be rewritten

$$E = \int \left(\tfrac{1}{2}d(\boldsymbol{x}, \boldsymbol{X}')^2 + \tfrac{1}{2}gh\right) h \, \mathrm{d}x \, \mathrm{d}y. \tag{35}$$

THEOREM 3. *Conditions for E in this form, and with constant f, to be stationary with respect to variations satisfying* (30) *within the domain D of the f-plane, and $\boldsymbol{n} \cdot \delta \boldsymbol{r} = 0$ on the boundary, are that $\boldsymbol{u} = \boldsymbol{u}_g$.*

Proof. The proof is just a rephrasing of theorem 2 in Cartesian coordinates, using (15) in particular, and the definitions of X' and Y' in (29). □

If we use (X, Y) to denote the stationary values of (X', Y'), the Cartesian version of the definitions (5) reappear as definitions

$$X = x + \frac{g}{f^2}\frac{\partial h}{\partial x}, \quad Y = y + \frac{g}{f^2}\frac{\partial h}{\partial y}, \tag{36}$$

in the present example of constant f. These X and Y are the geostrophic coordinates of Hoskins (1975) for shallow-water theory. They are called geostrophic coordinates because, when f is a constant, we can show that the equations for momentum balance in Cartesian coordinates can be re-written in the form $\dot{X} = u_g$, $\dot{Y} = v_g$.

The proof that E can be uniquely minimized is then carried out by showing that, given σ as a non-negative function of coordinates (X', Y'), there is a unique mapping from the coordinates (X', Y') to the physical coordinates (x, y) that minimizes E and satisfies (32). The condition that no fluid can enter or leave D across the boundaries is enforced by requiring this mapping to be from $\mathbb{R}^2$ into D. In the next section, we will show that there is a unique minimizer of E in the spherical case if the class of variations is written in the form (31).

3.2. *Generalized definitions of distance on the sphere*

The distance function defined by (34) is a Euclidean distance rescaled by the (constant) factor, f. This suggests that on the sphere we use a distance function based on the Riemannian distance on the sphere rescaled by the local Coriolis parameter. The effect is that the transformation is obtained by minimizing the energy under a constraint of the form (31), instead of under the constraints (10) and (11). The two forms of constraint are equivalent in the f-plane case, as discussed above. We will show in the next section that a state which minimizes the energy under one form of constraint also minimizes it under the other.

We start by considering how to measure the lengths of paths on the surface of the sphere. In the following, we simplify the notation by writing $\boldsymbol{X}$ (in place of the $\boldsymbol{X}'$ used above) as the generic second point acting as the end point of a path which starts at the point $\boldsymbol{x}$ on the sphere. Particular solution values of such $\boldsymbol{X}$, like (36) above, will be identified in the text when they are needed without necessarily introducing fresh notation.

Let $\boldsymbol{r}$ be the position vector, from a fixed origin O in three-dimensional Euclidean space, to a generic point on the surface of the sphere. Let a function $\boldsymbol{r}(s)$ of the distance s define a path $\boldsymbol{r} = \boldsymbol{r}(s)$ on the surface. Confine attention to part of the path of length l, so that $0 \leqslant s \leqslant l$, between end points

$$\boldsymbol{r}(0) = \boldsymbol{x} \ \text{(say)}, \quad \boldsymbol{r}(l) = \boldsymbol{X} \ \text{(say)}. \tag{37}$$

An increment of position $\mathrm{d}\boldsymbol{r}$ along the path can be expressed as an ordered pair of physical components (i.e. coefficients of local orthogonal unit vectors) which we write as

$$\mathrm{d}\boldsymbol{r} = (a\cos\phi\,\mathrm{d}\lambda, a\,\mathrm{d}\phi). \tag{38}$$

The local unit tangent to a piecewise smooth path described by functions $\lambda(s)$ and $\phi(s)$ is the vector

$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s} = a\left(\cos\phi\frac{\mathrm{d}\lambda}{\mathrm{d}s}, \frac{\mathrm{d}\phi}{\mathrm{d}s}\right). \tag{39}$$

Again the components on the right-hand side are the coefficients of local orthogonal unit vectors.

Equation (34) suggests that we next construct the integral of the Coriolis parameter along the finite segment of the path $\boldsymbol{r} = \boldsymbol{r}(s)$ defined above, between the end points (37). This is

$$A = \int_0^l f[\boldsymbol{r}(s)]\,\mathrm{d}s \tag{40}$$

between $\boldsymbol{x}$ and $\boldsymbol{X}$. We shall see in the next section that A has some features in common with the action integral found in classical mechanics. From (39) we deduce, from our hypotheses that the local unit vectors are orthogonal, that

$$\mathrm{d}s^2 = a^2(\cos^2\phi\,\mathrm{d}\lambda^2 + \mathrm{d}\phi^2), \tag{41}$$

so that A can be rewritten symbolically as

$$A = a\int_{\boldsymbol{x}}^{\boldsymbol{X}} f(\cos^2\phi\,\mathrm{d}\lambda^2 + \mathrm{d}\phi^2)^{1/2}. \tag{42}$$

This symbolic form highlights the presence of the end points (37) in a different way to (40). We will write L for the integrand of A/a in (42) for use in § 4. It is clear that L is bounded at each point on the path. In the special case of constant f,

$$A = lf. \tag{43}$$

The value of A in (40) clearly depends on the path chosen. We define $d(\boldsymbol{x}, \boldsymbol{X})$ to be the minimum value of $A(\boldsymbol{x}, \boldsymbol{X})$ over geometrically possible paths joining $\boldsymbol{x}$ to $\boldsymbol{X}$. These paths are geodesics on the sphere rescaled by the Coriolis parameter.

We note in passing that the general setting for a treatment of geometries defined by metrics of the form given in (42) is known as *conformal geometry* because multiplying metrics of the type (41) by the function $f(\phi)$ preserves angles and ratios of distances in the new geometries defined by (42), provided $f > 0$; see Sewell (2002, § 2) for further comment on the subject of conformal transformations, and §§ 4.1 and 4.3 of this paper for a discussion of how we treat $f \leqslant 0$.

3.3. *Differential properties of the distance function on the sphere*

In this subsection we establish certain differential properties of the path integral (42). Given a path $\boldsymbol{r} = \boldsymbol{r}(s)$, of length l, between end points $\boldsymbol{x}$ and $\boldsymbol{X}$ in (37), and the associated value of A, we wish to allow the path, and in particular its end points, to vary, and to examine the effect on A.

First, if only the length l varies, along the local tangent to the end point $\boldsymbol{X}$, with the starting point $\boldsymbol{x}$ held fixed, then

$$\frac{\mathrm{d}A}{\mathrm{d}l} = f(\boldsymbol{X}) \tag{44}$$

immediately from (40) (and (43) provides special examples). Since

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}l} \equiv \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}l} \tag{45}$$

from (37) is the unit tangent (and a particular value of (39)) at the end, we can construct a vector gradient

$$\frac{\mathrm{d}A}{\mathrm{d}\boldsymbol{X}} \equiv \frac{\mathrm{d}A}{\mathrm{d}l}\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}l} = f(\boldsymbol{X})\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}l} \tag{46}$$

of A at the end, for this particular variation, of l alone.

More generally, we now imagine that the direction of the local tangent at the end $\boldsymbol{X}$ is allowed to vary, as well as the length l of the curve. The curve becomes piecewise smooth there (instead of smooth as just above), but the same construction of the vector gradient (46), can be repeated, but using the new end tangent vector.

A different proof of this last conclusion can be constructed using Hamiltonians as follows. The vector gradient is first defined by specifying its components with respect to local orthogonal unit vectors as

$$\frac{\mathrm{d}A}{\mathrm{d}\boldsymbol{X}} \equiv \frac{1}{a}\left(\frac{1}{\cos\phi}\frac{\partial A}{\partial\lambda}, \frac{\partial A}{\partial\phi}\right) \tag{47}$$

where, on the right-hand side, the partial derivatives are with respect to the end values of λ and ϕ. Both λ and ϕ will be available almost everywhere on the curve to act as the local path parameter, as an alternative to s. The exceptional points will be where the path is locally parallel to a λ-coordinate curve, so that only λ is available there, and where the path is locally parallel to a ϕ-coordinate curve, so that only ϕ is available there. Without further loss of generality and because the metric is independent of λ, we shall choose ϕ as the path parameter and write $\dot{\lambda} = \mathrm{d}\lambda/\mathrm{d}\phi$ for brevity. The integrand of A/a can therefore be written as

$$L(\dot{\lambda}, \phi) = f(\phi)(1 + \cos^2\phi\dot{\lambda}^2)^{1/2}. \tag{48}$$

Equation (48) can be used as a Lagrangian to define a momentum $p = \partial L/\partial\dot{\lambda}$ and, via the standard Legendre transformation, a Hamiltonian $H(p, \phi) = p\dot{\lambda} - L$, such that $\dot{\lambda} = \partial H/\partial p$. (Here we are using standard extrapolation, in the calculus of variations, of terminology that originates in classical mechanics.) Away from the equator, so that $f \neq 0$, and away from the poles, so that $\cos\phi > 0$, we find that

$$H(p, \phi) = -\frac{(f^2\cos^2\phi - p^2)^{1/2}}{\cos\phi}. \tag{49}$$

The positive square root is chosen throughout these calculations. Then at each ϕ, the function $H(p)$ is strictly convex.

We can now write (42) as

$$\frac{A}{a} = \int p\mathrm{d}\lambda - \int H\mathrm{d}\phi \tag{50}$$

between limits of integration which are those values of λ and ϕ corresponding to the end points $\boldsymbol{x}$ and $\boldsymbol{X}$, i.e. to $s = 0$ and $s = L$. Differentiating with respect to those end values in (47) gives, at $\boldsymbol{X}$,

$$\frac{\partial A}{\partial \boldsymbol{X}} = \left(\frac{p}{\cos\phi}, -H\right). \tag{51}$$

We can write (39) at $\boldsymbol{X}$ as

$$\frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}s} = \frac{f}{L}(\cos\phi\dot{\lambda}, 1), \tag{52}$$

because $f\mathrm{d}s = aL\mathrm{d}\phi$, i.e. $\mathrm{d}\phi/\mathrm{d}s = f/aL$.

Equating components, we see that the first of (46) holds if and only if

$$\frac{p}{\cos\phi} = \frac{f^2\cos\phi\dot{\lambda}}{L}, \quad -H = \frac{f^2}{L}. \tag{53}$$

It is easy to verify that (53) is satisfied by using the properties $p = \partial H/\partial\dot{\lambda}$ and $H = p\dot{\lambda} - L$ with (48).

We can show, using similar techniques, that a similar result holds for $\partial A/\partial\boldsymbol{x}$:

$$\frac{\partial A}{\partial \boldsymbol{x}} = \left(-\frac{p}{\cos\phi}, H\right). \tag{54}$$

3.4. *Duality relations*

The new coordinates defined in (36) facilitate some very remarkable simplifications to solution strategies for the semi-geostrophic equations on an f-plane. These developments are reviewed and discussed in some depth in Sewell (2002) and McIntyre & Roulstone (2002, § 4). An important feature of the transformation between Lagrangian positions, $\boldsymbol{x}$, and the new variables, $\boldsymbol{X}$, is that the map is a Legendre transformation, in which isolated singularities can be interpreted as fronts (Sewell 2002; Chynoweth & Sewell 1989). A Legendre transformation arises in this case as a consequence of a duality between the two coordinate systems and two functions – the mass distribution $h(\boldsymbol{x}, t)$ and a stream function for the flow in $\boldsymbol{X}$-space, which we shall denote by $\mathbb{H}(\boldsymbol{X}, t)$ – that is defined by the following relationship

$$g\mathbb{H}(\boldsymbol{X}, t) - gh(\boldsymbol{x}, t) = \tfrac{1}{2}d^2(\boldsymbol{X}, \boldsymbol{x}), \tag{55}$$

where d is the rescaled Euclidean distance defined by (34) with $X' = X, Y' = Y$. We shall now proceed to establish certain relationships between derivatives of $\mathbb{H}$ and h, when d is now defined by (42) and following equations, and which reduce to those in f-plane theories when d is given by (34). In this section, our methods assume sufficient smoothness to allow derivatives etc. to exist in the usual sense. When formulating a procedure for solving our equations, we must deal with singularities of h and $\mathbb{H}$: we will address this issue in § 4. Referring to (55), noting that $d(\boldsymbol{X}, \boldsymbol{x})$ depends on time only through the time dependence of $\boldsymbol{X}$ and $\boldsymbol{x}$, we have

$$d\frac{\partial d}{\partial t} = 0 = g\frac{\partial \mathbb{H}}{\partial t} - g\frac{\partial h}{\partial t}. \tag{56}$$

This is a consequence of the passive variable nature of t in the duality expressed by (55).

By adopting ϕ as the path parameter, and then using (51), (54) and (53) with $A = d$, we have, from (55),

$$d\frac{\partial d}{\partial \boldsymbol{r}}\bigg|_{\boldsymbol{r}=\boldsymbol{X}} = g\frac{\partial \mathbb{H}}{\partial \boldsymbol{X}} = d\left(\frac{p}{\cos\phi}, \frac{f^2}{L}\right)\bigg|_{\boldsymbol{r}=\boldsymbol{X}}, \tag{57}$$

$$d\frac{\partial d}{\partial \boldsymbol{r}}\bigg|_{\boldsymbol{r}=\boldsymbol{x}} = -g\frac{\partial h}{\partial \boldsymbol{x}} = -d\left(\frac{p}{\cos\phi}, \frac{f^2}{L}\right)\bigg|_{\boldsymbol{r}=\boldsymbol{x}}. \tag{58}$$

Because the Lagrangian $L(\phi, \dot{\lambda})$ is independent of λ, then, by the Euler–Lagrange equation

$$\frac{\mathrm{d}}{\mathrm{d}\phi}\left(\frac{\partial L}{\partial \dot{\lambda}}\right) - \frac{\partial L}{\partial \lambda} = 0,$$

the 'momentum' $p \equiv \partial L/\partial \dot{\lambda}$ is constant along the path. Hence we can deduce the following relationships between the gradients of h and $\mathbb{H}$: from (51) and (54) (using the notation $\boldsymbol{X} = (\varLambda, \varPhi), \boldsymbol{x} = (\lambda, \phi)$)

$$\frac{g}{a\cos\varPhi}\frac{\partial \mathbb{H}}{\partial \varLambda} = \frac{pd}{\cos\varPhi}, \qquad \frac{g}{a\cos\phi}\frac{\partial h}{\partial \lambda} = \frac{pd}{\cos\phi},$$

and therefore we have (by cross-multiplication)

$$\frac{\partial \mathbb{H}}{\partial \varLambda} = \frac{\partial h}{\partial \lambda}, \tag{59}$$

and from

$$\frac{g}{a}\frac{\partial \mathbb{H}}{\partial \varPhi} = \frac{f(\boldsymbol{X})^2}{L(\boldsymbol{X})}d$$

and

$$\frac{g}{a}\frac{\partial h}{\partial \phi} = \frac{f(\boldsymbol{x})^2}{L(\boldsymbol{x})}d$$

(where the functional dependence of f and L on $\boldsymbol{x}$ and $\boldsymbol{X}$ means that these functions are evaluated at the respective end points) we have

$$\frac{f(\boldsymbol{x})^2}{L(\boldsymbol{x})}\frac{\partial \mathbb{H}}{\partial \varPhi} = \frac{f(\boldsymbol{X})^2}{L(\boldsymbol{X})}\frac{\partial h}{\partial \phi}. \tag{60}$$

Note that, when f is a constant and the path is a straight line (which means that $L(\boldsymbol{x}) = L(\boldsymbol{X})$)), then in Cartesian coordinates (59) and (60) reduce to $\partial \mathbb{H}/\partial \boldsymbol{X} = \partial h/\partial \boldsymbol{x}$, which is the shallow-water version of the gradient property of the geostrophic momentum transformation (cf. Hoskins 1975).

4. Solution of the semi-geostrophic equations on the sphere

4.1. *Definitions and notation*

In this section, we show that the coordinate transformation proposed in the previous section can be constructed, and exploit it to show that the semi-geostrophic equations can be solved on the sphere. We require a number of definitions to allow the problem to be stated and solved in a well-defined way.

Define a functional M as the obvious generalization of the energy as written in (35) for the f-plane case:

$$M = \tfrac{1}{2}\int_D (d(\boldsymbol{x},\boldsymbol{X})^2 + gh)h\,\mathrm{d}\Sigma. \tag{61}$$

We first show that the condition for M to be minimized, for a given $\sigma = h(\partial\boldsymbol{x}/\partial\boldsymbol{X})$, defines $\boldsymbol{x}$ implicitly as a function of $\boldsymbol{X}$. D denotes either the whole surface of the sphere or a bounded region of it, as in § 2.2. We will then show that the equation for the solution value of $\boldsymbol{X}$ in terms of $\boldsymbol{x}$ can be interpreted as defining a state of geostrophic balance, with energy equal to the minimizing value of M. We will thus have achieved a generalization of the geostrophic coordinate transformation.

We write the required mapping which determines $\boldsymbol{x}$ as a function of $\boldsymbol{X}$ as a mapping $\boldsymbol{s}$ from the surface of the sphere, $\mathscr{S}^2$, to itself. For this to be useful, we will require this mapping to be invertible, so that we can write $\boldsymbol{X} = \boldsymbol{s}^{-1}(\boldsymbol{x})$. Assume the point $\boldsymbol{X}$ has spherical coordinates (Λ, Φ) and the point $\boldsymbol{x}$ has coordinates (λ, ϕ). Following (32), we define σ to be

$$\sigma = h\frac{\partial(\lambda,\phi)}{\partial(\Lambda,\Phi)}\frac{\cos\phi}{\cos\Phi}, \tag{62}$$

so that σ is a given function of (Λ, Φ).

We now seek to minimize (61) for the given σ. We can write (61) as an integral over the transformed coordinates as

$$M = \tfrac{1}{2}\int_D (d(\boldsymbol{s}(\boldsymbol{X}),\boldsymbol{X})^2 + gh(\boldsymbol{s}(\boldsymbol{X})))\sigma(\boldsymbol{X})\,\mathrm{d}\nu, \tag{63}$$

where $\mathrm{d}\nu$ is the area measure

$$\mathrm{d}\nu = a^2\cos\Phi\,\mathrm{d}\Lambda\,\mathrm{d}\Phi. \tag{64}$$

The constraint of given σ is difficult to enforce. It can be made mathematically tractable using the concept of measure-preserving mappings. (See, for instance, Douglas (2002) for formal definitions.) Given a Borel set $B \subset \mathscr{S}^2$, define the measures

$$\left.\begin{aligned}\nu(B) &= \int_B \mathrm{d}\nu = \int_B a^2\cos\Phi\,\mathrm{d}\Lambda\,\mathrm{d}\Phi,\\ \varpi(B) &= \int_B a^2 h\cos\phi\,\mathrm{d}\lambda\,\mathrm{d}\phi.\end{aligned}\right\} \tag{65}$$

Thus $\varpi(B)$ measures the physical mass of fluid contained in B, and $\nu(B)$ measures the area of B in the transformed coordinates. We show that measure-preserving mappings are natural in this context. Calculate the image of each point on the sphere by setting $\boldsymbol{X} = \boldsymbol{s}^{-1}(\boldsymbol{x})$. Then, given a Borel set B, we can calculate $\varpi(\boldsymbol{x} : \boldsymbol{s}^{-1}(\boldsymbol{x}) \in B)$. The definition of σ yields that

$$\begin{aligned}\int_B \sigma\,\mathrm{d}\nu &= \int_B a^2 h\frac{\partial\boldsymbol{x}}{\partial\boldsymbol{X}}\frac{\cos\phi}{\cos\Phi}\cos\Phi\,\mathrm{d}\Lambda\,\mathrm{d}\Phi\\ &= \int_{(\boldsymbol{x}:\boldsymbol{s}^{-1}(\boldsymbol{x})\in B)} a^2 h\cos\phi\,\mathrm{d}\lambda\,\mathrm{d}\phi = \varpi(\boldsymbol{x} : \boldsymbol{s}^{-1}(\boldsymbol{x}) \in B).\end{aligned} \tag{66}$$

We define a measure $\sigma\nu$ by

$$\sigma\nu(B) = \int_B \sigma\,\mathrm{d}\nu \tag{67}$$

for all Borel sets B. Requiring that (66) holds for every Borel set B is exactly the definition that $\boldsymbol{s}^{-1}$ is a measure-preserving mapping from $\mathscr{S}^2$ (endowed with measure ϖ) to $\mathscr{S}^2$ (endowed with measure $\sigma\nu$). We will be concerned (initially) with the set S of all measure-preserving mappings from $(\mathscr{S}^2, \sigma\nu)$ to $(\mathscr{S}^2, \varpi)$; we will identify a unique energy minimizing measure-preserving mapping $\boldsymbol{s}$, which is invertible, and $\boldsymbol{s}^{-1}$ will satisfy property (66). We make the assumption (of non-degeneracy) that σ is ν-integrable, which means in particular that $\varpi(\boldsymbol{x} : \boldsymbol{s}^{-1}(\boldsymbol{x}) \in B) > 0$ implies $\nu(B) > 0$ for all sets B.

Minimizing (63) over the set of measure-preserving mappings S is an example of a general class of problems called *optimal mass transfer problems*; one seeks an optimal measure-preserving strategy which minimizes the 'transportation cost', where optimality is measured against a cost function. A review of these problems is given by Gangbo & McCann (1996). The integrand in (63) is an example of a cost function. This problem has a long history and has found many applications in physics, economics and statistics; the original problem posed by Gaspard Monge in 1781 was how to transport material between two locations in the most efficient way. McCann (2001) proves that this problem can be solved uniquely when the integrand in the cost function takes the form of the square of a distance function on a Riemannian manifold. The form of (63) suggests that the first term of the integrand could be interpreted in this way. However, we will need to extend McCann's results to deal with the whole of (63).

We mainly discuss the case where D is the whole spherical surface $\mathscr{S}^2$. The modifications to the argument where D is a bounded subset of it are discussed after the proof of theorem 5. The next step is therefore to define a manifold $\mathscr{M}$ to be $\mathscr{S}^2$ endowed with the distance function $d(\boldsymbol{X}, \boldsymbol{x})$ defined in the previous subsection. We will assume that d is a twice continuously differentiable function of $\boldsymbol{X}$ and $\boldsymbol{x}$. To apply McCann's theorem, we require $\mathscr{M}$ to be a compact, connected manifold without boundary. Since the Coriolis parameter f defines the distance function in (40), and f goes to zero at the equator, these assumptions will not be satisfied. We therefore regularize the problem, so that the assumptions are satisfied. Having solved the regularized problem, we will then have to show that a solution of the original problem can be recovered in the limit as the parameter defining the regularization tends to zero.

Define the Riemannian manifold $\mathscr{M}_\epsilon$ to be the surface of the sphere together with the metric $\hat{g}^{S^2}$ whose components take the form

$$\hat{g}_{ij}^{S^2} = \mathscr{F}^2(\phi) g_{ij}^{S^2}, \ \hat{g}^{ij^{S^2}} = \mathscr{F}^{-2}(\phi) g^{ij^{S^2}}, \tag{68}$$

where $g_{ij}^{S^2}$ denotes the usual components of the metric on a sphere S^2 of radius a,

$$g_{ij}^{S^2} = \begin{pmatrix} a^2 & 0 \\ 0 & a^2\cos^2\phi \end{pmatrix}. \tag{69}$$

$\mathscr{F}(\phi)$ is chosen to be a smooth modification of the function $2|\Omega \sin\phi|$ which is a twice differentiable function of ϕ, is equal to $2|\Omega \sin\phi|$ for $\phi > \phi_e > 0$ for some (small) ϕ_e and has a minimum value $\epsilon > 0$. Note that the metric would not change sign even without the regularization. The regularization is equivalent to using $f_\epsilon = \mathscr{F}(\phi(\boldsymbol{r}))$ in (40) to calculate A, and defining the distance function d by finding the minimum value of A. In §4.3, when the complete solution has been obtained, we will consider the effect of letting $\phi_e, \epsilon \to 0$. The resulting manifold $\mathscr{M}_\epsilon$ is topologically equivalent to the sphere.

Noting that we have assumed that σ is ν-integrable, use of McCann (2001, theorem 8) shows that

$$C(\boldsymbol{s}) = \tfrac{1}{2}\int_{\mathscr{M}_\epsilon} d^2(\boldsymbol{s}(\boldsymbol{X}), \boldsymbol{X})\sigma(\boldsymbol{X})\,\mathrm{d}\nu \tag{70}$$

has a unique minimizer $\boldsymbol{t}$ over $\boldsymbol{s} \in S$, where S is the set of all measure-preserving mappings from $(\mathscr{M}_\epsilon, \sigma\nu)$ to $(\mathscr{M}_\epsilon, \varpi)$, which can be expressed in the form

$$\boldsymbol{t}(\boldsymbol{X}) = \exp_{\boldsymbol{X}}[-\nabla\Psi(\boldsymbol{X})], \tag{71}$$

where Ψ is a scalar function and the gradient is taken with respect to the metric on $\mathscr{M}_\epsilon$. The operator on the right-hand side of (71) is the exponential map, as used by McCann (2001). The exponential map is a map from the tangent space of $\mathscr{M}_\epsilon$ at $\boldsymbol{X}$ to the manifold. The existence of the function Ψ characterizes optimality of the mapping $\boldsymbol{t}$ by specifying the direction, given by $-\nabla\Psi$, and the distance, given by $|\nabla\Psi|$, in which to move material from $\boldsymbol{X}$ to other locations in $\mathscr{M}_\epsilon$. The paths over which one moves material are the geodesics between points as defined by the metric on $\mathscr{M}_\epsilon$, and Ψ is determined up to an additive constant.

Moreover Ψ is involutive, a geometric condition which we explain below. Define a conjugate or dual function Ψ^c to Ψ by

$$\Psi^c(\boldsymbol{x}) = \inf_{\boldsymbol{X}}\left(\tfrac{1}{2}d^2(\boldsymbol{X}, \boldsymbol{x}) - \Psi(\boldsymbol{X})\right). \tag{72}$$

In general, a function φ need not satisfy $\varphi^{cc} = \varphi$ (where $\varphi^{cc} \equiv (\varphi^c)^c$); we call functions which do have this property involutive. A function is involutive exactly when it is the conjugate of some function (see Rachev & Ruschendorf 1998; §3.3; McCann 2001).

4.2. *Theorems and their consequences*

We now prove that the mapping $\boldsymbol{t}$ minimizes (63) under the constraint $\delta\sigma = 0$ using an argument similar to that in Cullen & Gangbo (2001) for the f-plane case. The strategy is as follows: we examine the energy functional when evaluated with the minimizer of (70) and show that perturbations always generate positive increments to this functional.

We first make some additional definitions. Write μ for the area measure in physical space, so that

$$\mu(B) = \int_B a^2\cos\phi\,\mathrm{d}\lambda\,\mathrm{d}\phi \tag{73}$$

and, for a μ-integrable function $\eta : \mathscr{M}_\epsilon \to \mathbb{R}$, write $\eta\mu$ for the measure defined by

$$\eta\mu(B) = \int_B a^2\eta(\lambda, \phi)\cos\phi\,\mathrm{d}\lambda\,\mathrm{d}\phi \tag{74}$$

for Borel sets $B \subset \mathscr{M}_\epsilon$. Thus, the measure ϖ can be written as $h\mu$.

Let a ν-integrable σ be given. We call a pair $(\boldsymbol{s}, \eta)$ *admissible* if $\boldsymbol{s}$ is an invertible measure-preserving mapping from $(\mathscr{M}_\epsilon, \sigma\nu)$ to $(\mathscr{M}_\epsilon, \eta\mu)$. We think of $\boldsymbol{s}^{-1}$ as a (possible) coordinate transformation, and η as a (possible) depth function. We write $\mathscr{S} \times \mathscr{H}$ for the set of admissible pairs. We will show that there is a unique pair $(\boldsymbol{t}, h) \in \mathscr{S} \times \mathscr{H}$ which minimizes

$$M(\boldsymbol{s}, \eta) = \tfrac{1}{2}\int_{\mathscr{M}_\epsilon} (d(\boldsymbol{s}(\boldsymbol{X}), \boldsymbol{X})^2 + g\eta(\boldsymbol{s}(\boldsymbol{X})))\sigma(\boldsymbol{X})\,\mathrm{d}\nu \tag{75}$$

over $(\boldsymbol{s}, \eta) \in \mathscr{S} \times \mathscr{H}$. This is equivalent to minimizing (63) integrated over $\mathscr{M}_\epsilon$ for given σ.

The problem of minimizing M cannot be rewritten as a standard mass transfer problem as solved by McCann (2001) as we have not fixed h. However, his theorem can be used in the proof. We begin by fixing η, and considering only the first part of the integrand in (75). This allows us to determine a $\boldsymbol{t}_\eta$ that depends on η. We then show that (75) can be minimized as a function of η. If h is the choice of η that achieves the minimization, then $(\boldsymbol{t}_h, h)$ solves the full problem. The statement that $-gh = \Psi^c$ is equivalent to saying that $-gh$ is involutive: this should be viewed as an extra constraint on the equations.

Theorem 4. *The integral* (75) *is uniquely minimized for* $(\boldsymbol{s}, \eta) \in \mathscr{S} \times \mathscr{H}$ *by* $(\boldsymbol{t}, h)$, *where* $\boldsymbol{t}$ *is the map* (71) *that minimizes* (70) *and* $-gh = \Psi^c$. Ψ^c *is defined by* (72), *using the* Ψ *that appears in* (71).

Proof. Start with any $\eta(\boldsymbol{x}) \geqslant 0$ such that $\int_{\mathscr{M}_\epsilon} \eta \,\mathrm{d}\mu = \int_{\mathscr{M}_\epsilon} \sigma \,\mathrm{d}\nu$. Use McCann's theorem to construct a map $\boldsymbol{t}_\eta$ which minimizes $C(\boldsymbol{s})$ for $\boldsymbol{s} : (\mathscr{M}_\epsilon, \sigma\nu) \to (\mathscr{M}_\epsilon, \eta\mu)$ as in (70). By construction, $(\boldsymbol{t}_\eta, \eta)$ is then an admissible pair. Use (75) to calculate $M(\boldsymbol{t}_\eta, \eta)$. Following arguments of Cullen & Gangbo (2001), this will be a strictly convex function of η (the second term is clearly strictly convex) and lower semicontinuous and coercive, and so can be uniquely minimized by some choice of η.

For any η, we can find a $\boldsymbol{t}$ and hence Ψ^c using (72) and (71). It is immediate from (72) that

$$\Psi(\boldsymbol{X}) + \Psi^c(\boldsymbol{x}) \leqslant \tfrac{1}{2} d^2(\boldsymbol{X}, \boldsymbol{x}), \tag{76}$$

for any $\boldsymbol{x}$, $\boldsymbol{X}$. McCann proves (in his theorem 7) that, if $\boldsymbol{t}$ is the minimizing map (71) with scalar function Ψ, then at every $\boldsymbol{X}$ where Ψ is differentiable, the inequality in (76) is strict unless $\boldsymbol{x} = \boldsymbol{t}(\boldsymbol{X})$, when it holds with equality. Thus

$$\Psi(\boldsymbol{X}) + \Psi^c(\boldsymbol{t}(\boldsymbol{X})) = \tfrac{1}{2} d^2(\boldsymbol{X}, \boldsymbol{t}(\boldsymbol{X})). \tag{77}$$

He also proves that Ψ is involutive. Using the non-degeneracy condition, he proves that $\boldsymbol{t}$ is invertible almost everywhere. It follows that $\boldsymbol{t}^{-1}(\boldsymbol{x}) = \exp_{\boldsymbol{x}}(-\nabla\Psi^c(\boldsymbol{x}))$.

We now make the 'guess' that the minimizer is characterized by setting $-g\eta = \Psi^c$, for the unique minimizer η as described in the first paragraph of the proof. This is consistent with the analysis of theorem 1. We justify this choice by noting that $(\exp_{\boldsymbol{X}}(-\nabla(-g\eta)^c(\boldsymbol{X})), \eta)$ is an admissible pair, and the fact that both $(-g\eta)^c$ and Ψ are involutive yields from McCann (2001, theorem 9) that $\boldsymbol{t}(\boldsymbol{X}) = \exp_{\boldsymbol{X}}(-\nabla(-g\eta)^c(\boldsymbol{X}))$ for (almost) every $\boldsymbol{X}$. For this choice, write $h = \eta$ and continue to write the minimizing map as $\boldsymbol{t}$. We demonstrate that this characterization indeed gives a minimizer as follows. Let $(\boldsymbol{s}, \eta)$ be an arbitrary member of $\mathscr{S} \times \mathscr{H}$. The definition (72) with $\boldsymbol{x}$ chosen to be $\boldsymbol{s}(\boldsymbol{X})$ gives

$$\Psi(\boldsymbol{X}) + \Psi^c(\boldsymbol{s}(\boldsymbol{X})) \leqslant \tfrac{1}{2} d^2(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{X})). \tag{78}$$

Now integrate (78) with respect to the measure $\sigma\nu$ to give

$$\int_{\mathscr{M}_\epsilon} \Psi(\boldsymbol{X})\sigma \,\mathrm{d}\nu + \int_{\mathscr{M}_\epsilon} \Psi^c(\boldsymbol{s}(\boldsymbol{X}))\sigma \,\mathrm{d}\nu \leqslant \int_{\mathscr{M}_\epsilon} \tfrac{1}{2} d^2(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{X}))\sigma \,\mathrm{d}\nu. \tag{79}$$

The inequality is strict if $\boldsymbol{s} \neq \boldsymbol{t}$. Now using the fact that $\boldsymbol{s} : (\mathscr{M}_\epsilon, \sigma\nu) \rightarrow (\mathscr{M}_\epsilon, \eta\mu)$ is measure-preserving, and identifying Ψ^c with $-gh$, we have

$$\int_{\mathscr{M}_\epsilon} \Psi(\boldsymbol{X})\sigma \, \mathrm{d}\nu - \int_{\mathscr{M}_\epsilon} gh\eta \, \mathrm{d}\mu \leqslant \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}d^2(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{X}))\sigma \, \mathrm{d}\nu, \tag{80}$$

with strict inequality if $\boldsymbol{s} \neq \boldsymbol{t}$. A similar calculation replacing $\boldsymbol{s}$ with $\boldsymbol{t}$ and η with h (and using (77)) gives

$$\int_{\mathscr{M}_\epsilon} \Psi(\boldsymbol{X})\sigma \, \mathrm{d}\nu - \int_{\mathscr{M}_\epsilon} gh^2 \, \mathrm{d}\mu = \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}d^2(\boldsymbol{X}, \boldsymbol{t}(\boldsymbol{X}))\sigma \, \mathrm{d}\nu. \tag{81}$$

Now

$$\begin{aligned} M(\boldsymbol{s}, \eta) - M(\boldsymbol{t}, h) = & \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}d^2(\boldsymbol{X}, \boldsymbol{s}(\boldsymbol{X}))\sigma \, \mathrm{d}\nu - \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}d^2(\boldsymbol{X}, \boldsymbol{t}(\boldsymbol{X}))\sigma \, \mathrm{d}\nu \\ & + \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}g\eta(\boldsymbol{s}(\boldsymbol{X}))\sigma \, \mathrm{d}\nu - \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}gh(\boldsymbol{t}(\boldsymbol{X}))\sigma \, \mathrm{d}\nu. \end{aligned} \tag{82}$$

The first integral in (82) is estimated using (80), and we use identity (81) to rewrite the second integral. In the third and fourth integrals we replace $\sigma \, \mathrm{d}\nu$ by $\eta \, \mathrm{d}\mu$ and $h \, \mathrm{d}\mu$, respectively (noting $\boldsymbol{s}$ and $\boldsymbol{t}$ are measure preserving). This gives

$$\begin{aligned} M(\boldsymbol{s}, \eta) - M(\boldsymbol{t}, h) & \geqslant g \int_{\mathscr{M}_\epsilon} (h^2 - \eta h) \, \mathrm{d}\mu + \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}g\eta^2 \, \mathrm{d}\mu - \int_{\mathscr{M}_\epsilon} \tfrac{1}{2}gh^2 \, \mathrm{d}\mu \\ & = \tfrac{1}{2}g \int_{\mathscr{M}_\epsilon} (h - \eta)^2 \, \mathrm{d}\mu. \end{aligned} \tag{83}$$

Thus, $M(\boldsymbol{s}, \eta) - M(\boldsymbol{t}, h) > 0$ unless η is equal to h; using the fact that the minimizer of (70) is unique, we deduce strict inequality unless $(\boldsymbol{s}, \eta)$ is equal to $(\boldsymbol{t}, h)$. The result follows. □

We can then deduce:

Theorem 5. *The integral* (61), *with* f_ϵ *as defined following* (69), *and* d *defined by the minimum of* A *in* (40), *is minimized with respect to displacements satisfying* (10) *and* $\delta\sigma = 0$ *if* $\boldsymbol{X}$ *is given as a function of* $\boldsymbol{x}$ *by the map*

$$\boldsymbol{X} = \exp_{\boldsymbol{x}}[g\nabla h(\boldsymbol{x})] \tag{84}$$

where $-gh(\boldsymbol{x})$ *is an involutive function. The minimizing value is*

$$\tfrac{1}{2} \int_{\mathscr{S}^2} \left(\boldsymbol{u}_g^{*2} + gh\right) h \, \mathrm{d}\mu, \tag{85}$$

where $\mathscr{F}\boldsymbol{u}_g^* = ((1/a\cos\phi)(\partial h/\partial\lambda), (1/a)(\partial h/\partial\phi))$. *In particular* $\boldsymbol{u}_g^* = \boldsymbol{u}_g$ *on those parts of* $\mathscr{M}_\epsilon$ *where* $\mathscr{F} = f$.

Proof. The definition of the space of admissible pairs $\mathscr{S} \times \mathscr{H}$ is consistent with $\delta\sigma = 0$. The definitions of the measures $\eta\mu$ and $h\mu$ are consistent with (10). The integral (61) takes the values $M(\boldsymbol{s}, \eta)$ given the map $\boldsymbol{s}(\boldsymbol{X})$, and the value $M(\boldsymbol{t}, h)$ given the map $\boldsymbol{t}(\boldsymbol{X})$. The proof of theorem 4 shows that $M(\boldsymbol{s}, \eta) - M(\boldsymbol{t}, h) > 0$ if $(\boldsymbol{s}, \eta) \neq (\boldsymbol{t}, h)$ and that the inverse map $\boldsymbol{t}^{-1}$ takes the form (84), as required.

Because of the identification of Ψ^c with $-gh$ made in the proof of theorem 4, and (77), we have that $-gh$ is involutive. Furthermore, the statement $\boldsymbol{t}^{-1}(\boldsymbol{x}) = \exp_{\boldsymbol{x}}[\nabla gh(\boldsymbol{x})]$ implies that the magnitude of $\boldsymbol{t}^{-1}(\boldsymbol{x})$ is a 'distance' $\mathscr{F}^{-1}\nabla(gh(\boldsymbol{x}))$ along the geodesic starting from $\boldsymbol{x}$. To see this, we note that $|g\nabla h| = (g^2 \hat{g}^{ij^{S^2}} \nabla_i h \nabla_j h)^{1/2} = (g^2 \mathscr{F}^{-2} \nabla^i h \nabla_i h)^{1/2}$ (using (68)). The minimizing value of $d^2(\boldsymbol{t}^{-1}(\boldsymbol{x}), \boldsymbol{x})$ is $|g\nabla h(\boldsymbol{x})|^2 = g^2 \mathscr{F}^{-2} \nabla^i h \nabla_i h = \boldsymbol{u}_g^{*2}$ as we require. □

We are now able to show that our analysis is a genuine extension of the constant Coriolis parameter energy minimization of Cullen & Gangbo (2001). If we consider the f-plane case, replacing $\mathscr{S}^2$ with some bounded region $\Omega \subset \mathbb{R}^2$, the geostrophic transformation is given by (36); if the minimum energy principle is satisfied we have

$$\boldsymbol{X} = (X, Y) = \nabla_e \left(\frac{|\boldsymbol{x}|_e^2}{2} + \frac{gh}{f^2} \right), \tag{86}$$

where $\boldsymbol{x} = (x, y)$, $|\cdot|_e$ denotes Euclidean distance, ∇_e denotes the usual derivative on $\mathbb{R}^2$, and $|\boldsymbol{x}|_e^2 + gh/f^2$ is a convex function at each time t. Our energy minimization result theorem 5 applied in the constant f setting yields $\boldsymbol{X}(\boldsymbol{x}) = \exp_{\boldsymbol{x}}(-\nabla \Psi^c(\boldsymbol{x}))$, where $\Psi^c = -gh$ is involutive. Thus,

$$\boldsymbol{X}(\boldsymbol{x}) = \exp_{\boldsymbol{x}}(\nabla gh(\boldsymbol{x})). \tag{87}$$

Recalling that the gradient in (87) is with respect to the rescaled metric, and that geodesics on $\mathbb{R}^2$ are straight lines so that the exponential map as characterized after (71) reduces to an elementary sum, we obtain

$$\boldsymbol{X}(\boldsymbol{x}) = \boldsymbol{x} + \nabla_e \left(\frac{gh}{f^2} \right) = \nabla_e \left(\frac{|\boldsymbol{x}|_e^2}{2} + \frac{gh}{f^2} \right). \tag{88}$$

Finally, given that a convex function may be characterized as the supremum of a family of continuous affine functions, we note that $-gh$ being involutive corresponds to convexity of $|\boldsymbol{x}|_e^2/2 + gh/f^2$. The advantage of working with involutive functions is that this is a concept that makes sense when generalized to a manifold, whereas global convexity need not; involutive functions enjoy analogous regularity properties to convex functions (see McCann 2001).

Further, we can now show how some of the issues concerning smoothness, mentioned in § 3.4, can be addressed using the analysis we have introduced here. Substituting $\Psi^c = -gh$, writing $\boldsymbol{t}(\boldsymbol{X}) = \boldsymbol{x}$, and defining $\mathbb{H} = (1/g)(-gh)^c$, we recover (55) from (77). Involutive functions are differentiable except on a set of zero size, therefore (57) and (58) are justified. In the f-plane case, the set of points where the convex function fails to be differentiable is thought to represent fronts: roughly speaking, this set has one dimension less than the domain. Analogous statements can be made about involutive functions, so the qualitative behaviour of the f-plane solutions is preserved on the manifold.

If the physical domain is a subset D of the surface of the sphere, as would arise in oceanographic applications, we can make the same definitions as above, including the use of the regularised Coriolis parameter $\mathscr{F}$. The support of σ will be some subset of $\mathscr{S}^2$, and we seek a mapping from $\mathscr{S}^2$ into D which minimizes (61) where the integral is taken over D. If the domain is small enough and far enough removed from the equator (so that all points are geodesically linked), the regularization will not be needed. In that case, theorem 13 of McCann (2001) can be used to show that an optimal map can be found.

Before we can use this result in the solution of the spherical semi-geostrophic equations, we have to resolve the issue that in § 2 we minimized the energy, subject to the constraints (10) and (11), while here we have used the constraints (10) and $\delta\sigma = 0$. It will be sufficient to prove that, if h is the depth function associated with an energy minimizer under one of these constraints, it is also the (depth) function associated with an energy minimizer under the other.

Theorem 6. *If h is the function that defines the minimizer of* (61) *as defined in Theorem 5, then there is a minimizer of* (61) *with respect to variations* (10) *and* (11), *defined by h via* (5) *and* (13). *Conversely, suppose we have a minimizer of* (61) *subject to the variations* (10) *and* (11). *Then the h associated with this minimizer defines a minimizer of* (61) *with respect to* (10) *and $\delta\sigma = 0$ as described in theorem* 5. *Moreover $-gh$ is involutive.*

Proof. We first show that the minimizer of (61) subject to variations satisfying (10) and $\delta\sigma = 0$ is also a minimizer of (61) under variations (10) and (11). A variation satisfying $\delta\sigma = 0$ can be characterised by saying that for each $\boldsymbol{X}$, the associated $\boldsymbol{x}$ changes in such a way that (10) is satisfied. As in § 2.3, we define a notional velocity $\boldsymbol{u}$. This is chosen to have magnitude equal to the distance from $\boldsymbol{x}$ to $\boldsymbol{X}$ along the geodesic connecting them, and direction normal to that geodesic. Thus if $\boldsymbol{x}$ and $\boldsymbol{X}$ are related by the minimizing map (84), we have $\boldsymbol{u} = \boldsymbol{u}_g$ as in theorem 2. Varying $\boldsymbol{x}$ along the connecting geodesic, with $\boldsymbol{X}$ fixed, yields

$$\delta\boldsymbol{u} = \delta\boldsymbol{x}^{\perp},$$

where $\perp$ indicates a rotation of a vector by a right angle. Writing this variation on the original sphere S^2 gives

$$\delta\boldsymbol{u} = f\delta\boldsymbol{x}^{\perp}.$$

Now consider variations of $\boldsymbol{x}$ normal to the geodesic. These will not change the magnitude of $\boldsymbol{u}$, but rotate it. Without loss of generality, we can then write the effect of a general variation of $\boldsymbol{x}$, using spherical polar coordinates as used in (11), as

$$\delta u = fa\delta\phi + v(\sin\phi\delta\lambda + \xi), \qquad \delta v = -fa\cos\phi\delta\lambda - u(\sin\phi\delta\lambda + \xi). \tag{89}$$

The scalar ξ depends on $\boldsymbol{x}$ and cannot be explicitly determined. The terms involving ξ show that the amount of rotation may not be the same as that given by (11).

The change to (61) is thus the same whether the variations are given by (11) or by $\delta\sigma = 0$. So, given a map $\boldsymbol{X}$ to $\boldsymbol{x}$ minimizing (61) under the conditions of theorem 5, apply a perturbation to it governed by (11). Following (89), the variation can be written as the combination of a variation satisfying $\delta\sigma = 0$, and a rotation of $\boldsymbol{u}$ which depends on $\boldsymbol{x}$. By assumption, the former makes a non-negative change to (61). The latter does not change (61). Thus the overall change is non-negative, and so the map $\boldsymbol{X}$ to $\boldsymbol{x}$ generates a state which minimizes (61) under the conditions of theorem 2.

Conversely, we have a minimizer of (61) under the variations (10) and (11). A minimizer is a stationary point, therefore (13) is satisfied. Compute σ from h by constructing $\boldsymbol{X}$ at each $\boldsymbol{x}$ using (84). Given any variation with $\delta\sigma = 0$, we can write the effect on u and v in the form (89). Thus the effect of the variation will be a combination of a variation satisfying (11) and a rotation of $\boldsymbol{u}$ which depends on $\boldsymbol{x}$. The change to the energy will again be non-negative. Therefore we have a minimizer under variations with $\delta\sigma = 0$, and so h must define the unique solution of the optimal map for the associated σ guaranteed by McCann's theorem. In particular $-gh$ is involutive. □

Note, that while it makes sense to seek a unique minimizer of (61) for a given σ, the minimization of (61) under variations (10) and (11) is a purely local property of h.

4.3. *Solution of the semi-geostrophic equations*

Cullen & Gangbo (2001) solve the shallow-water semi-geostrophic equations in the f-plane case by showing that they can be written as a transport equation in geostrophic coordinates, and proving that the transport equation can be solved. Though we will show in the next section that the equations can also be written as a transport equation in geostrophic coordinates on the sphere, the transport velocity can no longer be written down explicitly, so cannot be used directly as the basis of a solution procedure. We therefore have to solve the equations in physical space by showing that the energy minimization problem can be solved, generating a depth field h. We then exploit theorem 6 to show that $-gh$ is involutive and therefore has the same regularity properties as the convex functions used in the f-plane case. In particular, these properties guarantee that a sequence of approximations converges.

The first step is to apply the regularization discussed in §4.1 after (68) to the physical space equations. We can thus regard f as bounded away from zero. Semi-geostrophic solutions are invariant to a transformation which leaves h fixed, reverses the signs of f, $\partial/\partial\phi$, v and v_g, and leaves u, u_g and $\partial/\partial\lambda$ fixed. Thus we solve the equations with the sign of f reversed in the southern hemisphere, and then with f replaced with the strictly positive function $\mathscr{F}$ defined in §4.1. After finding a solution h_ϵ of the regularized problem, we take the limit as $\epsilon \to 0$. If the limit solution satisfies necessary continuity conditions at the equator, it can be transformed back to a solution of the original equations (3), (5) and (6).

The difficulty in finding energy minimizers subject to (11) is that these variations are non-integrable, as shown by Roulstone & Sewell (1997 equation (7.20)). Thus a finite displacement satisfying (11) does not give a well-defined change to $\boldsymbol{u}$. For instance, given $u = v = 0$ at $(0, 0)$, displace a particle at that position to the point $(\pi/4, \pi/4)$ and calculate the change in $\boldsymbol{u}$ according to (11). If the displacement proceeds via the point $(0, \pi/4)$, the result is $(\Omega a, -\Omega a(1 + \pi/4))$. If it is via $(\pi/4, 0))$, the result is $(\Omega a\sqrt{2}, 0)$.

We resolve this difficulty by defining a specific search direction, chosen to be a steepest descent path in energy. Assume we are given initial data for h and the notional velocity $\boldsymbol{u}$ such that

$$\chi = \left(v - \frac{g}{fa\cos\phi}\frac{\partial h}{\partial\lambda}, -u - \frac{g}{fa}\frac{\partial h}{\partial\phi}\right) \neq 0. \tag{90}$$

Minimize the energy (9) iteratively by calculating a displacement

$$\delta\boldsymbol{r} = \kappa f^{-1}\chi \tag{91a}$$

$$= \kappa(f^{-1}(v - v_g), -f^{-1}(u - u_g)), \tag{91b}$$

where $\boldsymbol{u}_g$ is calculated from h using (5), and using (10) and (11) to update h, u, v. κ is an iteration parameter. Substituting (91*b*) into (11) gives that

$$u\delta u + v\delta v = -\kappa(u(u - u_g) + v(v - v_g)). \tag{92}$$

Then using (5), (14) and (91), and assuming no displacements across any boundary, we obtain

$$\delta E = -\kappa\int((u - u_g)^2 + (v - v_g)^2)h\,\mathrm{d}\Sigma = -\kappa\int\chi^2 h\,\mathrm{d}\Sigma. \tag{93}$$

This is negative definite and vanishes when $\boldsymbol{u} = \boldsymbol{u}_g$. Since the energy is a positive definite quantity, the energy found by this iteration is bounded below, and convergence is thus guaranteed. According to theorem 6, the resulting state will be also a minimum of the energy under the constraint (31).

In order to use this in the semi-geostrophic solution procedure, we require the stronger property that the total reduction in energy is $O(\chi^2)$, at least for sufficiently small χ. To show this, use (90) and (91) to write

$$\int \chi \cdot \delta\chi h \, \mathrm{d}\Sigma = \int f\kappa^{-1}\delta\boldsymbol{r} \cdot \delta\chi h \, \mathrm{d}\Sigma. \tag{94}$$

Then use (11) to write the right-hand side as

$$\int f\kappa^{-1}\delta\boldsymbol{r} \cdot [-((fa\cos\phi\delta\lambda + u_g\sin\phi\delta\lambda,\ fa\delta\phi + v_g\sin\phi\delta\lambda) + \delta(v_g, -u_g))]h \, \mathrm{d}\Sigma. \tag{95}$$

The same manipulations that lead from (20) to (23) then give

$$\int \kappa^{-1}(\delta\boldsymbol{r} \cdot \boldsymbol{P} \cdot \delta\boldsymbol{r} + g(\nabla \cdot (h\delta\boldsymbol{r}))^2)h \, \mathrm{d}\Sigma$$
$$= -\int \kappa(f^{-2}\chi \cdot \boldsymbol{P} \cdot \chi + g(\nabla \cdot (hf^{-1}\chi))^2)h \, \mathrm{d}\Sigma, \tag{96}$$

where $\boldsymbol{P}$ is the matrix defined in (24). This is negative definite. We require the stronger condition that the right-hand side is $O(\chi^2)$. A proof of this is outside the scope of this paper. However, the only case where both terms of (96) vanish is where $f + (u_g \tan\phi)/a = 0$, $v_g = 0$ and χ is parallel to $\boldsymbol{u}_g$. This case corresponds to an anticyclonic vortex with zero semi-geostrophic absolute vorticity centred at the pole. The associated σ is a Dirac mass, and was excluded from the analysis of Cullen & Gangbo (2001) by assumptions on the initial data. In the present case, it will be necessary to prove that such a case cannot be generated from an appropriate choice of initial data. Provided that the right-hand side of (96) can be shown to be $O(\chi^2)$, the reduction of energy during the iteration will also be $O(\chi^2)$.

Assuming that our claim above can be proved, we can solve the regularized semi-geostrophic system as follows:

(i) Start with an initial $h_\epsilon(0)$ such that $-gh_\epsilon(0)$ is involutive. Calculate $\boldsymbol{u}_g$ from it.

(ii) Take a time step δt. Make a first guess at the solution by rotating $\boldsymbol{u}_g$ by an angle $f\delta t$ at each point.

(iii) Use the rotated $\boldsymbol{u}_g$ as the first guess notional velocity and minimize the energy under variations (10) and (11) using the procedure described above. The initial χ for the iteration will thus be $O(\delta t)$, and the energy will be reduced by $O(\delta t^2)$. The result will be $h_\epsilon(\delta t)$.

(iv) Use theorem 6 to show that $-gh_\epsilon(\delta t)$ is involutive.

(v) We now use the methods of Cullen & Gangbo (2001). For a given finite time interval, we discretize in time. For each choice of time step δt, we can generate a depth field $h_\epsilon(t)$ such that $-gh_\epsilon$ is involutive at each time. As the time step converges to zero, we generate a sequence of involutive functions. Given such a sequence, it is easy to see that they have a global Lipschitz bound. Now the Ascoli–Arzela theorem (on families of equicontinuous functions) yields the existence of a limit function. Moreover, standard arguments in the literature show that the limit function is involutive. The limit solution will conserve energy, as the total energy loss in the approximate solution over a fixed time interval will be $O(\delta t)$.

We now let ϵ tend to zero. The involutive condition on $-gh_\epsilon$ means that h must satisfy the conditions derived in (26) and the following text. These imply that v_g must tend to zero at $\phi = 0$ as ϵ tends to zero. We can therefore reverse the transformation used to regularize the problem, thus reversing the signs of f, $\partial/\partial\phi$, v and v_g, without creating any discontinuity.

5. Properties of the semi-geostrophic solution

5.1. *Transport equation in geostrophic coordinates*

We first show that the evolution equations reduce to a transport equation in geostrophic coordinates, as in the f-plane case. Thus the solution can be interpreted in terms of potential density advection and inversion.

We first use the exponential map, (84), to find $\boldsymbol{X}$ as a function of $\boldsymbol{x}$, $h(\boldsymbol{x}, t)$ and the derivatives of $h(\boldsymbol{x}, t)$. To illustrate the general method, we begin by showing how (84) leads to the usual expression (36) for the dual coordinates on an f-plane.

The dual coordinates, (36), can be expressed in the form

$$X_i = \exp\left(\epsilon \frac{\mathrm{d}}{\mathrm{d}s}\right) x_i, \tag{97}$$

where $X_i = (X, Y)$, $x_i = (x, y)$, s parameterizes the path between the two end points $(\boldsymbol{X}, \boldsymbol{x})$, and ϵ is a distance along the path (see, for example, Schutz 1980 §§ 2.13 and 5.3). To see this, we note that on the f-plane the path between $\boldsymbol{x}$ and $\boldsymbol{X}$ is a straight line, with metric given by (34), and a simple relationship can be derived, using the results of § 3.3, relating the gradient of d to the tangent vectors at either end point:

$$\frac{\mathrm{d}x_i}{\mathrm{d}s} = -\frac{1}{f}\frac{\partial d}{\partial x_i}, \quad \frac{\mathrm{d}X_i}{\mathrm{d}s} = \frac{1}{f}\frac{\partial d}{\partial X_i}. \tag{98}$$

The derivative $\mathrm{d}/\mathrm{d}s$ defines a vector field $\mathrm{d}/\mathrm{d}s = (\mathrm{d}x_i/\mathrm{d}s)\partial/\partial x_i = (\mathrm{d}X_i/\mathrm{d}s)\partial/\partial X_i$, and therefore (97) becomes

$$X_i = \left(1 + \epsilon \frac{\mathrm{d}x_j}{\mathrm{d}s}\frac{\partial}{\partial x_j} + \frac{\epsilon^2}{2!}\left(\frac{\mathrm{d}x_j}{\mathrm{d}s}\frac{\partial}{\partial x_j}\right)\left(\frac{\mathrm{d}x_k}{\mathrm{d}s}\frac{\partial}{\partial x_k}\right) + \cdots\right) x_i. \tag{99}$$

Because the path is a straight line, the tangent vector $\mathrm{d}x_i/\mathrm{d}s$ is constant along the path and therefore (99) becomes

$$X_i = x_i + \epsilon \frac{\mathrm{d}x_i}{\mathrm{d}s}. \tag{100}$$

Then, using (98) and (34), (100) becomes

$$X_i = x_i + \frac{\epsilon g}{fd}\frac{\partial h}{\partial x_i} \tag{101}$$

and if $\epsilon = d/f$ (cf. (43)), we recover (36).

In the case of spherical geometry and variable f, (98) becomes

$$\frac{\mathrm{d}x_i}{\mathrm{d}s} = -\frac{1}{f(\boldsymbol{x})}\frac{\partial d}{\partial x_i}, \quad \frac{\mathrm{d}X_i}{\mathrm{d}s} = \frac{1}{f(\boldsymbol{X})}\frac{\partial d}{\partial X_i}, \tag{102}$$

and we can substitute the first set of these relations into (99), whereupon, using (58), (99) becomes

$$X_i = x_i + \frac{\epsilon g}{df(\boldsymbol{x})}\frac{\partial h}{\partial x_i} + \frac{\epsilon^2}{2!}\left(\frac{g}{df(\boldsymbol{x})}\frac{\partial h}{\partial x_j}\frac{\partial}{\partial x_j}\right)\left(\frac{1}{df(\boldsymbol{x})}\frac{\partial h}{\partial x_i}\right) + \cdots. \tag{103}$$

Noting that $(g/f(\boldsymbol{x}))(\partial h/\partial x_i) = (v_g, -u_g) = -\boldsymbol{k}\times\boldsymbol{u}_g$, (103) can be written in the form

$$\boldsymbol{X} = \boldsymbol{x} - \frac{\epsilon}{d}\boldsymbol{k}\times\boldsymbol{u}_g + \frac{\epsilon^2}{2!}\frac{1}{d}((\boldsymbol{k}\times\boldsymbol{u}_g)\cdot\nabla)\left(\frac{\boldsymbol{k}\times\boldsymbol{u}_g}{d}\right) + \cdots. \tag{104}$$

We calculate $\mathrm{d}\boldsymbol{X}/\mathrm{d}t$ from (104), noting that the second-order term involves only changes to d and changes to $\boldsymbol{u}_g$ parallel to the curve connecting $\boldsymbol{x}$ and $\boldsymbol{X}$.

The first step is to differentiate (55) with respect to time following particles. This gives

$$g\dot{\mathbb{H}}(\boldsymbol{X},t) - g\dot{h}(\boldsymbol{x},t) = \dot{d}d(\boldsymbol{X},\boldsymbol{x}). \tag{105}$$

Writing $\dot{h} = \partial h/\partial t + \dot{\boldsymbol{x}}\cdot\nabla h$ and a similar expression for $\dot{\mathbb{H}}$ and using (56) gives

$$g\dot{\boldsymbol{X}}\cdot\nabla\mathbb{H} - g\dot{\boldsymbol{x}}\cdot\nabla h = \dot{d}d(\boldsymbol{X},\boldsymbol{x}). \tag{106}$$

In addition, we have from theorem 5 that $d = |\boldsymbol{u}_g^*|$. Thus we can rewrite (106) as

$$g\dot{\boldsymbol{X}}\cdot\nabla\mathbb{H} - g\dot{\boldsymbol{x}}\cdot\nabla h = \dot{\boldsymbol{u}}_g^*\cdot\boldsymbol{u}_g^*. \tag{107}$$

The semi-geostrophic equations (6) can be written in vector form as $\dot{\boldsymbol{u}}_g = \boldsymbol{k}\times(f\boldsymbol{u}_g - f\dot{\boldsymbol{x}})$, where $\boldsymbol{k}$ is a unit vector in the vertical. Assuming that the same equations control the evolution of the regularized variable $\boldsymbol{u}_g^*$ and substituting this into (107) gives

$$g\dot{\boldsymbol{X}}\cdot\nabla\mathbb{H} - g\dot{\boldsymbol{x}}\cdot\nabla h = f\dot{\boldsymbol{x}}\times\boldsymbol{u}_g^* = -g\dot{\boldsymbol{x}}\cdot\nabla h. \tag{108}$$

Thus we have $g\dot{\boldsymbol{X}}\cdot\nabla\mathbb{H} = 0$, so that $\dot{\boldsymbol{X}}$ is parallel to contours of $\mathbb{H}$. The magnitude of $\dot{\boldsymbol{X}}$ has to be determined by differentiating (104), and can only be determined when $\dot{\boldsymbol{x}}$ and $\dot{\boldsymbol{u}}_g^*$ are known.

In the f-plane theory, $\dot{\boldsymbol{X}} = \boldsymbol{u}_g$, which corresponds to using only the first two terms on the right-hand side of (104). It can then be shown that $\dot{\boldsymbol{X}}$ is non-divergent in $\boldsymbol{X}$ coordinates, and thus that potential density is conserved in a Lagrangian sense. In the present case, we can write

$$\dot{\boldsymbol{X}} = \left(-\frac{\chi}{f(\Phi)}\frac{\partial\mathbb{H}}{\partial\Phi}, \frac{\chi}{f(\Phi)\cos\Phi}\frac{\partial\mathbb{H}}{\partial\Lambda}\right).$$

This is the geostrophic wind calculated at $\boldsymbol{X}$ using the 'depth' field $\mathbb{H}$ multiplied by a factor χ which comes from the additional terms on the right-hand side of (104). These measure the curvature of the geodesic between $\boldsymbol{x}$ and $\boldsymbol{X}$ which arises from both the curvature of the original sphere, and the additional curvature of $\mathscr{M}_\epsilon$ induced by the conformal rescaling. χ is of order $1 + O(|\boldsymbol{u}_g|/fa)^2$, since curvature effects will only be significant if $\boldsymbol{x}$ and $\boldsymbol{X}$ are separated by a distance comparable to the earth's radius. For geostrophic winds of order $15\,\mathrm{m\,s^{-1}}$ as used in our example solutions, $|\boldsymbol{u}_g|/fa = 0.025$ in middle latitudes.

The divergence of $\dot{\boldsymbol{X}}$ comes from the variations in χ discussed above and from that of the geostrophic wind itself. We can remove the latter divergence by making the additional rescaling

$$\sin\Phi\,\mathrm{d}\tilde{\Lambda} = \mathrm{d}\Lambda, \quad \mathrm{d}\tilde{\Phi} = \mathrm{d}\Phi.$$

This changes the factor $\cos\Phi$ in the metric of the sphere to $\sin 2\Phi$, and transforms the sphere into two spheres, tangent at the equator. The use of this rescaling to create non-divergence was noted by Salmon (1985). The effect is that the scaled potential density

$$\tilde{\sigma} = \frac{h\partial(\lambda\cos\phi,\phi)}{f\partial(\Lambda\cos\Phi,\Phi)}$$

is almost conserved in a Lagrangian sense (subject to the curvature effects estimated in the previous paragraph). The inverse of this can readily be seen to be an approximate form of the Ertel potential vorticity.

Since σ is a measure of mass in phase space, standard kinematics yields the conservation law

$$\frac{\partial \sigma}{\partial t} = \nabla \cdot (\sigma \dot{\boldsymbol{X}}). \tag{109}$$

Equations (104) and (109) define the evolution in $\boldsymbol{X}$ coordinates. The continuity equation (2) implies that, within any material circuit defined by fixed values of the Lagrangian coordinates α and β:

$$\int h(\alpha, \beta, 0)\, \mathrm{d}\alpha\, \mathrm{d}\beta = a \int h(\alpha, \beta, t)\, \mathrm{d}\mu = a \int \sigma(\alpha, \beta, t)\, \mathrm{d}\nu. \tag{110}$$

This takes the form of a conservation of 'circulation' in phase space, though it has nothing to do with the circulation in physical space. It is a semi-geostrophic analogy of the 'impermeability' result of Haynes & McIntyre (1990).

5.2. *Calculation of the coordinate transformation by energy minimization*

In this subsection, we use the iteration defined in (91) to calculate the coordinate transformation for typical meteorological data. We then show some solutions, and illustrate the use of potential density as a diagnostic. We use the model of Mawson (1996). Start with a given $\sigma(\lambda, \Phi)$ defined by (62). Choose a first guess solution for h and the notional velocity $\boldsymbol{u}$

$$h(\Lambda, \Phi) = \sigma h_0, u = 0, v = 0. \tag{111}$$

We now construct a displacement $\delta \boldsymbol{r}$ as defined by (91) so that the energy is minimized under (10) and (11). If the displacement takes each point (Λ, Φ) to a point (λ, ϕ), then conservation of mass as expressed by (10) implies that

$$\sigma = h \frac{\partial(\lambda, \phi)}{\partial(\Lambda, \Phi)} \frac{\cos \phi}{\cos \Phi}. \tag{112}$$

We claim that, by making this special choice of initial data, we have constructed a solution of the minimization problem with $\delta\sigma = 0$. The length of the paths between each $\boldsymbol{x}$ and $\boldsymbol{X}$ on the sphere is given by integrating $\delta \boldsymbol{r}$ from $\boldsymbol{x}$ to $\boldsymbol{X}$. Since we have used (11) to update u and v, and the final u and v equal their geostrophic values, the value of the action integral (40) along the path is equal to $|\boldsymbol{u}_g|$. By construction, the iteration procedure preserves σ and satisfies (10), the energy minimizing state it reaches must be the one found in theorem 5. The iteration paths are exactly the length of the geodesics mapping points $\boldsymbol{x}$ to $\boldsymbol{X}$ in theorem 5, so they must be the same as those geodesics.

As before, the choice $\sigma/h_0 = 1$ represents a trivial state of balance with no flow. An example of a non-trivial choice is shown in figure 1(*a*). This is designed to reproduce disturbances typical of a low-level atmospheric pressure field with geostrophic winds of about $15\,\mathrm{m\,s^{-1}}$.

The solution defines a map from points $\boldsymbol{X}$ with coordinates (Λ, Φ) to points $\boldsymbol{x}$ with coordinates (λ, ϕ) by summing the displacements over all the iterations to give $(\delta\lambda_t, \delta\phi_t)$, and setting $\lambda = \Lambda + \delta\lambda_t, \phi = \Phi + \delta\phi_t$.

The result of applying the procedure to the first guess field shown in figure 1 is illustrated in figure 2. A hundred iterations were used. The semi-geostrophic shallow-water model of Mawson (1996) was used. The initialization procedure described there

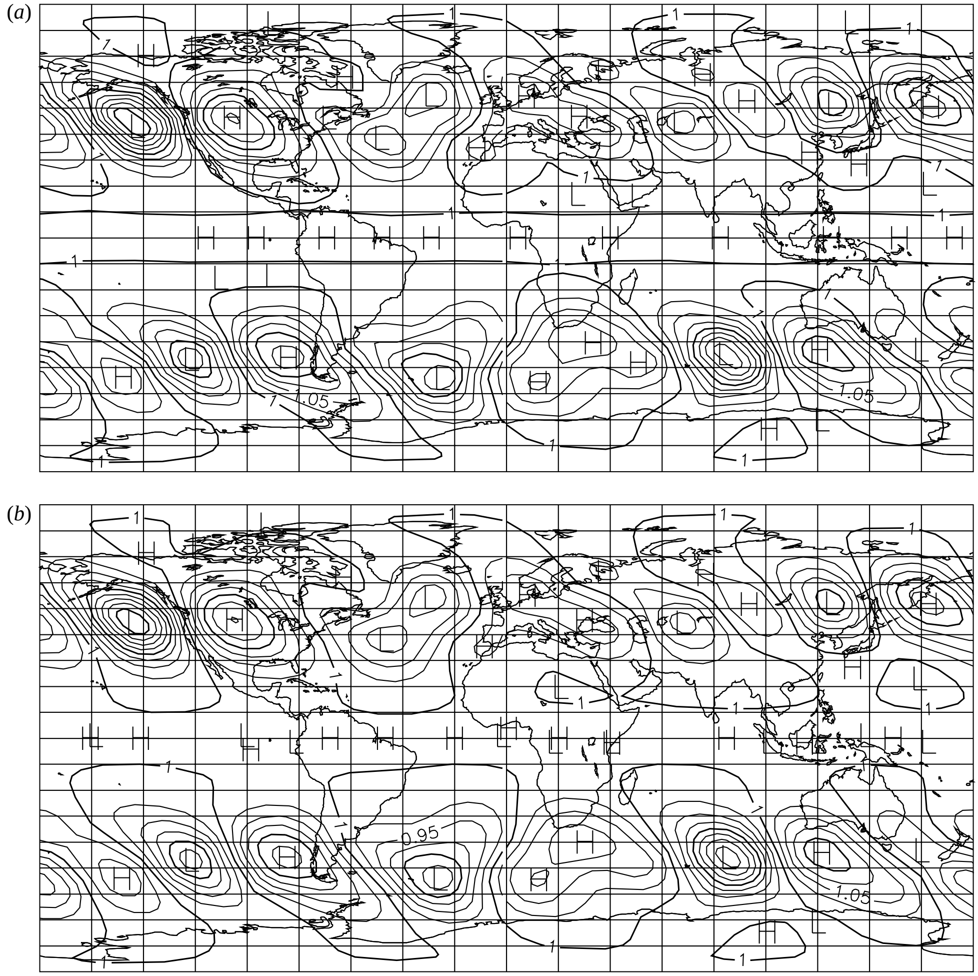

FIGURE 1. (*a*) Initial distribution of the dimensionless quantity σ/h_0 for equation (111). Contour interval 0.025. (*b*) Final distribution of σ/h_0. Contour interval 0.025.

(p. 276: initialization stage 2) is equivalent to using (91). The 'correction velocity' U_A defined on p. 271 of that paper generates the displacement required by (91), and the updates using (10) and (11) are equivalent to equations (10) and (9) in Mawson (1996). Figure 2 shows that positive anomalies in σ/h_0 become positive anomalies of h. The h field is smoother than the σ field. This is to be expected, since σ is related to the potential vorticity, which is expected to have smaller scales than the depth field.

As a check, the procedure can be reversed. Given initial data with depth h and initial winds $u = u_g$, $v = v_g$, choose $\delta\mathbf{r}$ as minus the value given by (91), and iterate to a state where $u = v = 0$. Set σ equal to the final value of h. This procedure acts as a diagnosis of potential density from a given geostrophically balanced state. Since the initial magnitude of $\delta\mathbf{r}$ given by (91) will be zero, we must use a semi-implicit procedure of the form

$$\delta\mathbf{r}_n = -\tfrac{1}{2}\kappa\left(\left(fv - \frac{1}{a\cos\phi}\frac{\partial h}{\partial\lambda}, -fu - \frac{1}{a}\frac{\partial h}{\partial\phi}\right)_{n-1} + \left(fv - \frac{1}{a\cos\phi}\frac{\partial h}{\partial\lambda}, -fu - \frac{1}{a}\frac{\partial h}{\partial\phi}\right)_n\right), \tag{113}$$

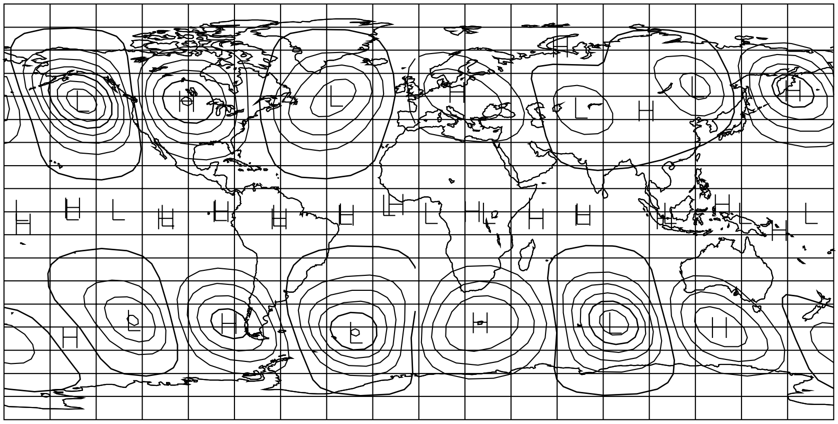

FIGURE 2. Distribution of h derived from initial field shown in figure 1. Contour interval 250 m. Bold contours at 10 600, 10 700, 10 800 m.

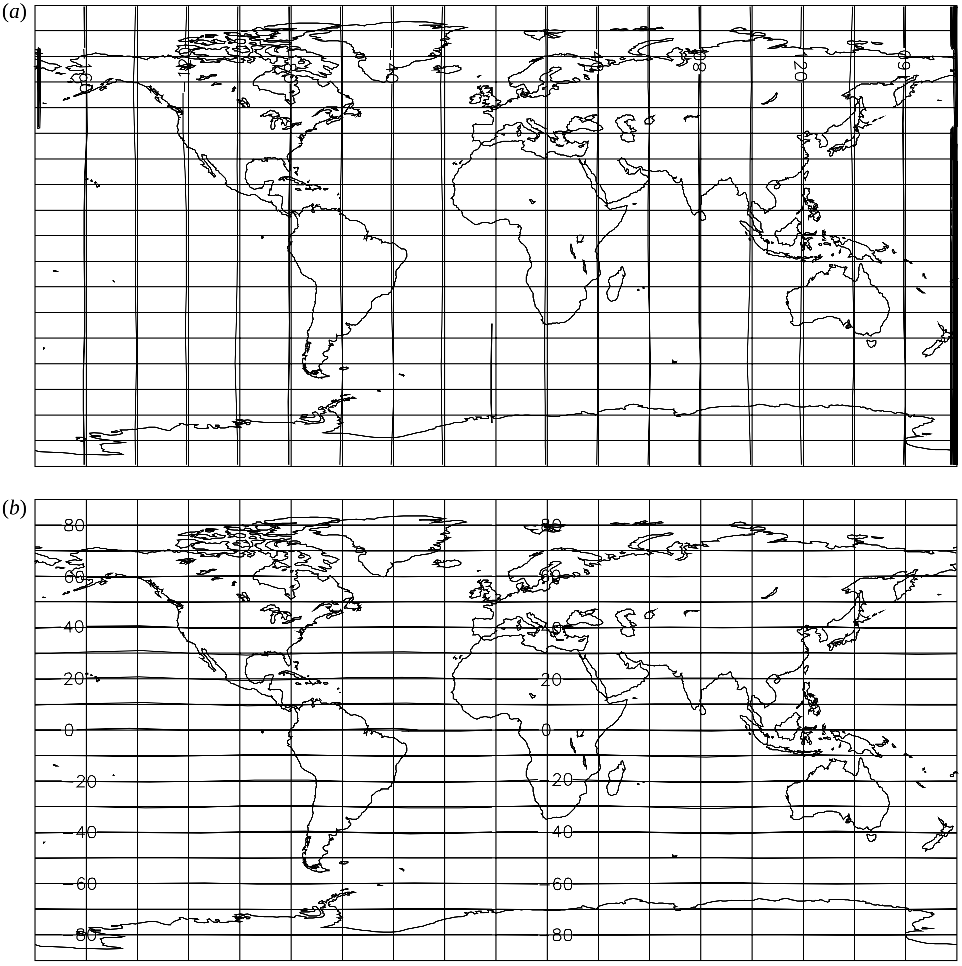

FIGURE 3. (a) Λ plotted as a function of λ, ϕ, contours every 20°. (b) Φ plotted against λ, ϕ. Contours every 10°.

Initial σ	Minimizing h	Final σ
12 324	10 826	12 313
8384	10 533	8358

TABLE 1. Test of potential density inversion procedure.

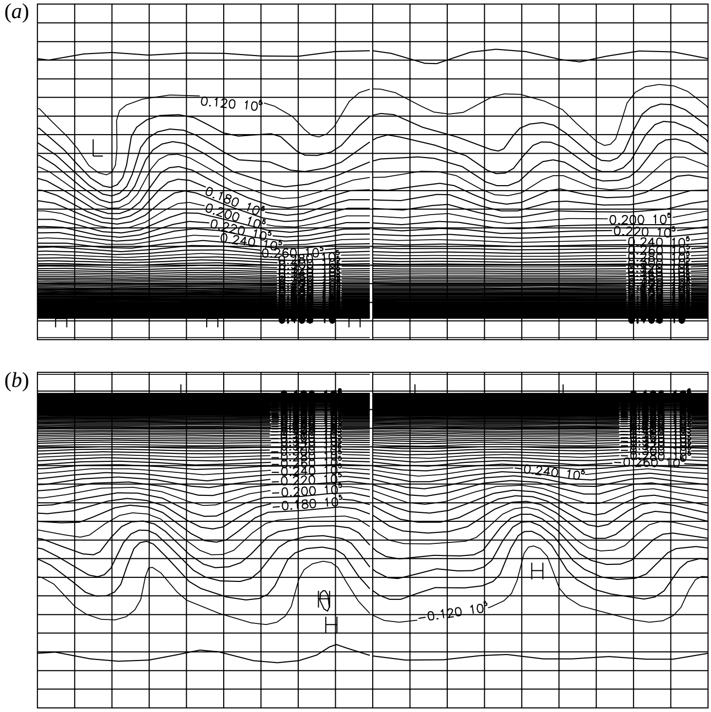

FIGURE 4. Distribution of $\tilde{\sigma}/h_0$ derived from data shown in figure 1. Contour interval 10^4 s. (*a*) Northern hemisphere. (*b*) Southern hemisphere.

where (10) and (11) with $\delta \boldsymbol{r} = \delta \boldsymbol{r}_n$ are used to update h, u and v from values at iteration level $n - 1$ to values at iteration level n. The final values of h will be equal to the original σ, subject to numerical error. Figure 1(*b*) illustrates the final field. It is almost identical to the original field. Figure 3 shows values of (Λ, Φ) calculated from (λ, ϕ) by summing the displacements defined in (113) over all the iterations. For the data chosen, the displacements are quite small, and the displacement of the latitude and longitude grid lines is only just visible. (Λ, Φ) can be regarded as the natural generalization of geostrophic coordinates on the sphere; the coordinate transformation is not far from the identity for these data.

The maximum and minimum values of h are also set out in table 1. They show that the calculation of h from σ has been reversed to within about 1 %. Tests with reducing κ and increasing the number of iterations show convergence of the error to zero. The errors come both from the early relatively large iteration steps and from accumulated numerical error over all the iterations. Finite iteration steps only exactly follow the steepest descent path in the limit $\kappa \to 0$.

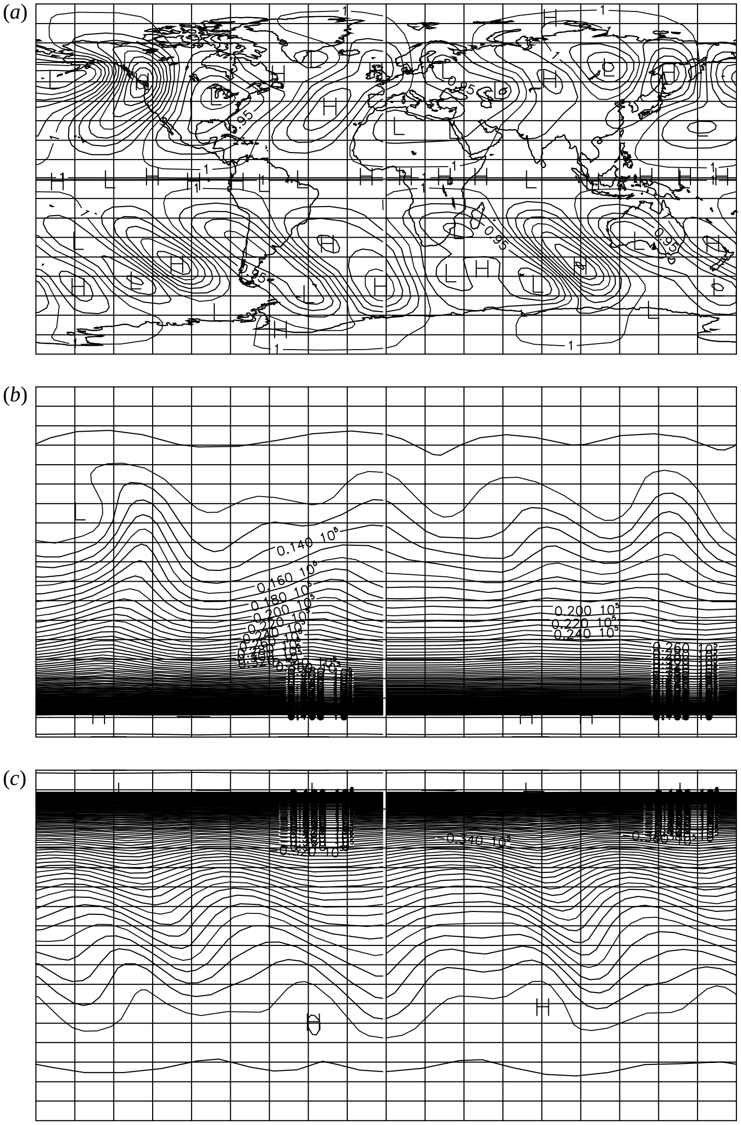

FIGURE 5. (*a*) Distribution of $\sigma/(h_0)$ after a two-day forecast from initial data shown in figure 1. Contour interval 0.025. (*b*) Distribution of $\tilde{\sigma}/h_0$ at the same time (Northern hemisphere). Contour interval 10^4 s. (*c*) As (*b*), but for Southern hemisphere.

We finally show an example of the evolution of the semi-geostrophic model using the initial conditions illustrated in figure 1. We plot the approximately conserved quantity $\tilde{\sigma}/h_0$. Figure 4 shows the initial data corresponding to figure 1. Values close to the equator are not plotted. The L_2 norm of $\tilde{\sigma}/h_0$ is infinite, so instead we calculate the L_2 norm of the determinant of the matrix (24) divided by fh. This is the physical space form of the semi-geostrophic potential vorticity, calculated using the local value of f.

The fields of σ/h_0 and $\tilde{\sigma}/(h_0)$ are shown in figures 5 and 6 after 2 and 20 days integration, respectively. The L_2 norm of the physical space potential vorticity is increased by a factor of 1.0006 after 2 days, and by 1.0029 after 20 days. Since the

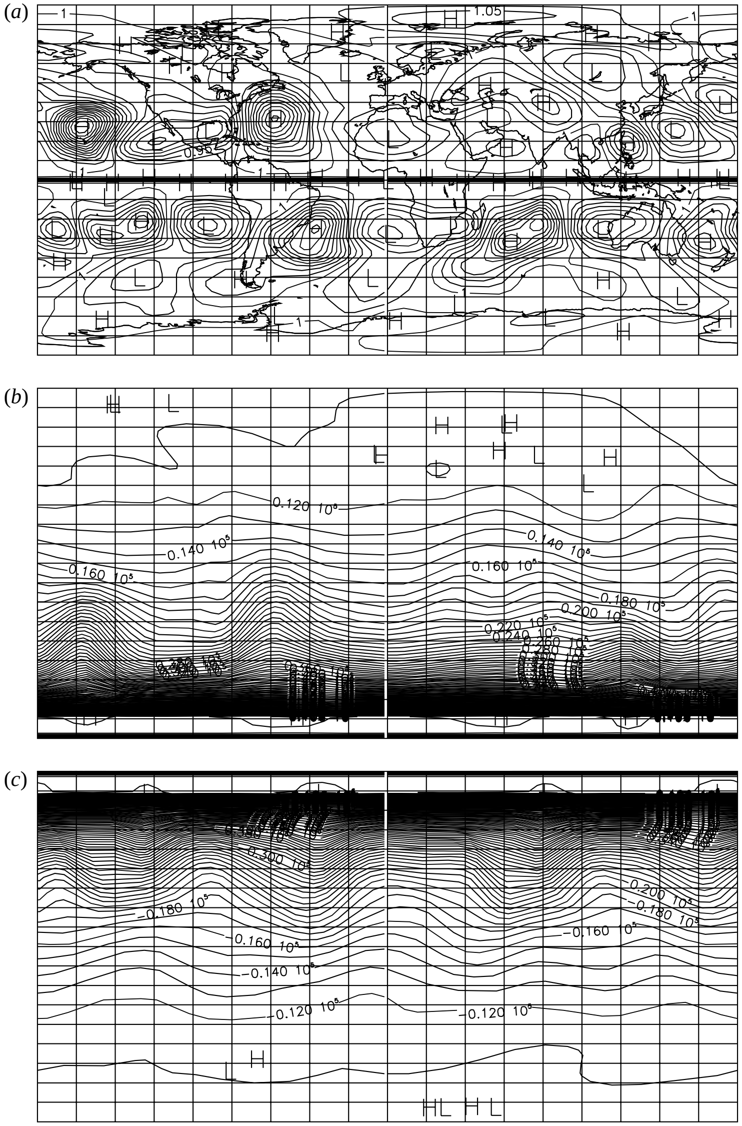

FIGURE 6. (*a*) Distribution of $\sigma/(h_0)$ after 20-day forecast from initial data shown in figure 1. Contour interval 0.025. (*b*) Distribution of $\tilde{\sigma}/h_0$ at the same time (Northern hemisphere). Contour interval 10^4 s. (*c*) As (*b*), but for Southern hemisphere.

numerical methods used do not conserve potential vorticity exactly in the f-plane case, it would be difficult to determine whether potential vorticity is conserved or not from this diagnostic. The worst-case estimate made in the previous subsection is almost certainly an overestimate of the effect on a global integral. Comparison of figures 3 and 5 shows that the disturbances propagate faster near the equator, as expected from the dispersion formula for Rossby waves in spherical semi-geostrophic theory which is identical to that derived from the primitive shallow-water equations, (Mawson 1996, p. 280). After 20 days, the disturbances have migrated closer to the equator, thus providing a severe test for the integration scheme.

6. Issues for further work

We have shown how the geostrophic coordinate transformation can be extended to the sphere in a well-defined way. The convexity principle which makes the coordinate transformation well-defined in the f-plane case has a natural generalization. If a potential vorticity or density is defined, respectively, as the mass-weighted Jacobian of the forward and reverse transformation, we have shown how a potential density inversion procedure can be implemented. The evolution equations take the form of the transport of potential density by a 'velocity' parallel to the geostrophic wind. We have shown formally how the resulting equations can be solved.

There are several remaining analytic issues to be resolved in order to make the above solution procedure rigorous. The energy minimizing argument of Cullen & Gangbo (2001) has to be extended to Riemannian manifolds, and the proof that the physical space energy minimization problem can be solved with an energy reduction of $O(\chi^2)$ has to be made rigorous. In addition, we have to prove that the limit of the regularized semi-geostrophic solutions on the sphere is well-defined as the regularization parameter ϵ tends to zero. Since the proofs that solutions exist rely on the convergence of a sequence of approximations to the depth field, and the depth field is very flat near the equator, this is likely to be straightforward.

Meteorologically, the constraints on the dynamical system resulting from the conservation laws need to be explored, in particular, to establish which flows are nonlinearly stable under this type of dynamics. An important issue is the apparent lack of an equivalent to the absolute vorticity conservation law satisfied by the barotropic vorticity equation on the sphere. This may be related to the need to enforce inertial stability. Inertial stability is not required to make the barotropic vorticity equation soluble. The inertial stability condition prevents the model describing genuinely two-dimensional disturbances to the depth field near the equator. This may reflect the clearly different dynamics that is observed near the equator, for instance the inability of tropical cyclones to form close to the equator.

The first version of this work was carried out at the Isaac Newton Institute programme on 'Atmosphere-Ocean Dynamics' held from July to December 1996. The main part of the work was carried out under EPSRC grant no. GR/L43220. MJPC wishes to thank Robert McCann, Mikhail Feldman and Bo Su who provided useful input during a visit to Madison, Wisconsin, funded by NSF grant no. DMS 00 74 037 and also Cedric Villani who provided useful input during a visit to Reading in January 2003. We would also like to thank the anonymous referees for their constructive input.

Appendix

Theorem 7. *Let $\mathscr{S}$ be a curved surface lying in a three-dimensional space, having unit normal $\boldsymbol{N}$, and bounded by a closed curve $\mathscr{C}$ whose unit normal locally tangent to the surface is $\boldsymbol{n}$. Let $\boldsymbol{B}$ be a vector field which, on $\mathscr{S}$, is tangent to $\mathscr{S}$. Then*

$$\int_{\mathscr{S}} [\mathrm{div}\boldsymbol{B} - \boldsymbol{N}\cdot(\boldsymbol{N}\cdot\mathrm{grad})\boldsymbol{B}]\,\mathrm{d}S = \oint_{\mathscr{C}} \boldsymbol{n}\cdot\boldsymbol{B}\,\mathrm{d}s, \tag{A 1}$$

where ds *is the measure of distance along $\mathscr{C}$, and* div *and* grad *are the vector differential operators in three-dimensional space.*

Proof. This is a corollary of Stokes' theorem, which states that

$$\int_{\mathscr{S}} \boldsymbol{N}\cdot(\mathrm{curl}\boldsymbol{A})\,\mathrm{d}S = \oint_{\mathscr{C}} \boldsymbol{A}\cdot\boldsymbol{t}\,\mathrm{d}s \tag{A 2}$$

for any three-dimensional smooth vector field $\boldsymbol{A}$, where $\boldsymbol{t}$ is the local unit tangent to $\mathscr{C}$ and curl is the vector differential operator in three dimensions.

First choose any smooth vector field $\boldsymbol{B}$, and construct $\boldsymbol{A}=\boldsymbol{N}\times\boldsymbol{B}$ on $\mathscr{S}$. Then on $\mathscr{C}$

$$\boldsymbol{t}\cdot\boldsymbol{A}=\boldsymbol{t}\cdot(\boldsymbol{N}\times\boldsymbol{B})=(\boldsymbol{t}\times\boldsymbol{N})\cdot\boldsymbol{B}=\boldsymbol{n}\cdot\boldsymbol{B},$$

by defining $\boldsymbol{n}=\boldsymbol{t}\times\boldsymbol{N}$. Hence,

$$\oint_{\mathscr{C}} \boldsymbol{A}\cdot\boldsymbol{t}\,\mathrm{d}s = \oint_{\mathscr{C}} \boldsymbol{n}\cdot\boldsymbol{B}\,\mathrm{d}s$$

which is the required form of the right-hand side of (A 1). Second invoke the vector analysis identity

$$\mathrm{curl}(\boldsymbol{N}\times\boldsymbol{B})=(\boldsymbol{B}\cdot\mathrm{grad})\boldsymbol{N}-\boldsymbol{B}(\mathrm{div}\boldsymbol{N})-(\boldsymbol{N}\cdot\mathrm{grad})\boldsymbol{B}+\boldsymbol{N}(\mathrm{div}\boldsymbol{B}),$$

then on $\mathscr{S}$

$$\boldsymbol{N}\cdot[\mathrm{curl}(\boldsymbol{N}\times\boldsymbol{B})]=\boldsymbol{N}\cdot(\boldsymbol{B}\cdot\mathrm{grad})\boldsymbol{N}-\boldsymbol{N}\cdot\boldsymbol{B}(\mathrm{div}\boldsymbol{N})-\boldsymbol{N}\cdot(\boldsymbol{N}\cdot\mathrm{grad})\boldsymbol{B}+\mathrm{div}\boldsymbol{B}$$

because $\boldsymbol{N}\cdot\boldsymbol{N}=1$. Now invoke the properties that $\boldsymbol{B}$ is tangential to the surface $\mathscr{S}$, so that $\boldsymbol{N}\cdot\boldsymbol{B}=0$, and

$$\boldsymbol{N}\cdot(\boldsymbol{B}\cdot\mathrm{grad})\boldsymbol{N}=\boldsymbol{N}\cdot\frac{\partial\boldsymbol{N}}{\partial B}=\tfrac{1}{2}\frac{\partial\boldsymbol{N}\cdot\boldsymbol{N}}{\partial B}=0,$$

because $\boldsymbol{N}\cdot\boldsymbol{N}=1$ and $\boldsymbol{B}\cdot\mathrm{grad}=\partial/\partial B$ is the derivative in the direction of $\boldsymbol{B}$. Thus,

$$\int_{\mathscr{S}}[\mathrm{div}\boldsymbol{B}-\boldsymbol{N}\cdot(\boldsymbol{N}\cdot\mathrm{grad})\boldsymbol{B}]\,\mathrm{d}S=\oint_{\mathscr{C}}\boldsymbol{n}\cdot\boldsymbol{B}\,\mathrm{d}s \tag{A 3}$$

and because the differential operators are defined in three-dimensional space, $\boldsymbol{N}(\boldsymbol{N}\cdot\mathrm{grad})$ is the gradient in the $\boldsymbol{N}$-direction, therefore grad $-\ \boldsymbol{N}(\boldsymbol{N}\cdot\mathrm{grad})$ is the gradient tangential to the surface. □

The application of this result to (8) requires $\boldsymbol{B}=h^2\dot{\boldsymbol{r}}$ and then we obtain the required result

$$\int_{\mathscr{S}}[\mathrm{div}(h^2\dot{\boldsymbol{r}})-\boldsymbol{N}\cdot(\boldsymbol{N}\cdot\mathrm{grad})h^2\dot{\boldsymbol{r}}]\,\mathrm{d}S=\oint_{\mathscr{C}}h^2\boldsymbol{n}\cdot\dot{\boldsymbol{r}}\,\mathrm{d}s.$$

REFERENCES

Benamou, J.-D. & Brenier, Y. 1998 Weak existence for the semi-geostrophic equations formulated as a coupled Monge–Ampère/transport problem. *SIAM J. Appl. Maths* **58**, 1450–1461.

Chynoweth, S. & Sewell, M. J. 1989 Dual variables in semi-geostrophic theory. *Proc. R. Soc. Lond.* A **424**, 155–186.

Cullen, M. J. P. 2000 On the accuracy of the semi-geostrophic approximation. *Q. J. R. Met. Soc.* **126**, 1099–1116.

Cullen, M. J. P. & Douglas, R. J. 1998 Applications of the Monge–Ampère equation and Monge transport problem to meteorology and oceanography. *Contemporary Maths* **226**, 33–53.

Cullen, M. J. P. & Gangbo, W. 2001 A variational approach for the 2-D semi-geostrophic shallow water equations. *Arch. Rat. Mech. Anal.* **156**, 241–273.

Cullen, M. J. P. & Purser, R. J. 1984 An extended Lagrangian theory of semi-geostrophic frontogenesis. *J. Atmos. Sci.* **41**, 1477–1497.

Cullen, M. J. P. & Purser, R. J. 1989 Properties of the Lagrangian semi-geostrophic equations. *J. Atmos. Sci.* **46**, 2684–2697.

Cullen, M. J. P., Norbury, J., Purser, R. J. & Shutts, G. J. 1987 Modelling the quasi-equilibrium dynamics of the atmosphere. *Q. J. R. Met. Soc.* **113**, 735–757.

Douglas, R. J. 1998 Rearrangements of vector valued functions, with application to atmospheric and oceanic flows. *SIAM J. Math. Anal.* **29**, 891–902.

Douglas, R. J. 2002 Rearrangements of functions, with application to meteorology and ideal fluid flow. *Large-Scale Atmosphere-Ocean Dynamics: Vol. 1 Analytic Methods & Numerical Models* (ed. J. Norbury & I. Roulstone), pp. 288–341. Cambridge University Press.

Eliassen, A. 1948 The quasi-static equations of motion. *Geofys. Publ.* **17**, no. 3.

Gangbo, W. & McCann, R. J. 1996 The geometry of optimal transportation. *Acta Math.* **177**, 113–161.

Haynes, P. H. & McIntyre, M. E. 1990 On the conservation and impermeability theorems for potential vorticity. *J. Atmos. Sci.* **47**, 2021–2031.

Hoskins, B. J. 1975 The geostrophic momentum approximation and the semi-geostrophic equations. *J. Atmos. Sci.* **32**, 233–242.

Hoskins, B. J., McIntyre, M. E. & Robertson, A. W. 1985 On the use and significance of isentropic potential vorticity maps. *Q. J. R. Met. Soc.* **111**, 887–946.

McCann, R. J. 2001 Polar factorization of maps on Riemannian manifolds. *Geomet. Funct. Anal.* **11**, 589–608.

McIntyre, M. E. & Roulstone, I. 2002 Are there higher-accuracy analogues of semi-geostrophic theory? *Large-Scale Atmosphere-Ocean Dynamics: Vol. 2 Geometric Methods and Models* (ed. J. Norbury & I. Roulstone), pp. 301–364. Cambridge University Press.

Magnusdottir, G. & Schubert, W. H. 1991 Semi-geostrophic theory on the hemisphere. *J. Atmos. Sci.* **48**, 1449–1456.

Mawson, M. H. 1996 A shallow water semi-geostrophic model on a sphere. *Q. J. R. Met. Soc.* **122**, 267–290.

Mawson, M. H. & Cullen, M. J. P. 1992 An idealised simulation of the Indian monsoon using primitive-equation and quasi-equilibrium models. *Q. J. R. Met. Soc.* **118**, 153–164.

Phillips, N. A. 1963 Geostrophic motion. *Rev. Geophys.* **1**, 123–176.

Polvani, L. M. & Sobel, A. H. 2002 The Hadley circulation and the weak temperature gradient approximation. *J. Atmos. Sci.* **59**, 1744–1752.

Purser, R. J. 1999 Legendre-transformable semi-geostrophic theories. *J. Atmos. Sci.* **56**, 2522–2535.

Rachev, S. T. & Ruschendorf, L. 1998 *Mass Transportation Problems*. Springer.

Roulstone, I. & Sewell, M. J. 1996 Potential vorticities in semi-geostrophic theory. *Q. J. R. Met. Soc.* **122**, 983–992.

Roulstone, I. & Sewell, M. J. 1997 The mathematical structure of theories of semi-geostrophic type. *Phil. Trans. R. Soc. Lond.* A 2489–2517.

Salmon, R. 1985 New equations for nearly geostrophic flow. *J. Fluid Mech.* **153**, 461–477.

Schubert, W. H. 1985 Semi-geostrophic theory. *J. Atmos. Sci.* **42**, 1770–1772.

Schubert, W. H., Ciesieleski, P. E., Stevens, D. E. & Kuo, H.-C. 1991 Potential vorticity modelling of the ITCZ and the Hadley circulation. *J. Atmos. Sci.* **48**, 1493–1509.

Schutz, B. 1980 *Geometrical Methods of Mathematical Physics*. Cambridge University Press.

Sewell, M. J. 2002 Some applications of transformation theory in mechanics. *Large-Scale Atmosphere-Ocean Dynamics: Vol. 2 Geometric Methods and Models* (ed. J. Norbury & I. Roulstone), pp. 143–223. Cambridge University Press.

Shutts, G. J. 1989 Planetary semi-geostrophic equations. *J. Fluid Mech.* **208**, 545–573.

Shutts, G. J. & Cullen, M. J. P. 1987 Parcel stability and its relation to semigeostrophic theory. *J. Atmos. Sci.* **44**, 1318–1330.