Balanced data assimilation for highly oscillatory mechanical systems

Gottfried Hastermann* Maria Reinhardt† Rupert Klein‡ Sebastian Reich§

October 21, 2020

Abstract

Data assimilation algorithms are used to estimate the states of a dynamical system using partial and noisy observations. The ensemble Kalman filter has become a popular data assimilation scheme due to its simplicity and robustness for a wide range of application areas. Nevertheless, the ensemble Kalman filter also has limitations due to its inherent Gaussian and linearity assumptions. These limitations can manifest themselves in dynamically inconsistent state estimates. We investigate this issue in this paper for highly oscillatory Hamiltonian systems with a dynamical behavior which satisfies certain balance relations. We first demonstrate that the standard ensemble Kalman filter can lead to estimates which do not satisfy those balance relations, ultimately leading to filter divergence. We also propose two remedies for this phenomenon in terms of blended time-stepping schemes and minimization based post-processing methods. The effects of these modifications to the standard ensemble Kalman filter are discussed and demonstrated numerically for the balanced motion of highly oscillatory Hamiltonian systems. In our context this scenario serves as prototypical example for applications from meteorology.

Keywords. Data assimilation, ensemble Kalman filter, balanced dynamics, highly oscillatory systems, Hamiltonian dynamics, geophysics
AMS (MOS) subject classifications. 65C05, 62M20, 93E11, 62F15, 86A22

*Freie Universität Berlin, Institut für Mathematik, Arnimallee 6, D-14195 Berlin, Germany
†Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany
‡Freie Universität Berlin, Institut für Mathematik, Arnimallee 6, D-14195 Berlin, Germany
§Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany

1 Introduction

A problem dating back as far as the advent of numerical weather prediction is the incorporation of physical observations into a dynamical model with more than one time scale. The famous first forecast of L. F. Richardson [33] spectacularly failed, due to the choice of an unbalanced initial condition gained from observations. In essence the observational data did not satisfy certain discrete energy balances which triggered the artificial oscillations in the pressure, ultimately leading to the wrong result. In the context of data assimilation procedures, several solutions to the issue of artificial fast oscillations in dynamical systems with multiple time scales were proposed over the last decades. In [34] it was suggested to apply a digital filter after every assimilation step to filter out spurious fast oscillations. This was applied to a weather prediction model. Strategies that incorporate the observational data in the model evolution in a gradual and smooth way instead of using all the information about the observation at one single point in time have been suggested for example in [7] and [6]. Other methods to overcome the issue of artificial balances triggered by Bayesian data assimilation were proposed in [26] and [21]. In the context of variational data assimilation the issue was addressed e.g. in [13].

1.1 Model problem

With atmospheric models in mind and in line with the seminal investigations of [30, 31] and, more specifically, [8, 32, 11] we will address the issue of representing systems with multiple time scales that evolve in approximate balance with respect to their fast modes in a simplified finite dimensional setting. In the following we therefore discuss sequential data assimilation techniques for highly-oscillatory Hamiltonian systems with Hamiltonian energy functional

$$H^\varepsilon(q,p) = \frac{1}{2}p^{\mathrm{T}}p + \frac{1}{2\varepsilon^2}g(q)^{\mathrm{T}}Kg(q) + V(q), \tag{1}$$

with momenta and coordinates $p, q \in \mathbb{R}^N$. Here $V : \mathbb{R}^N \to \mathbb{R}$ is a potential, $g : \mathbb{R}^N \to \mathbb{R}^L$, $L \leq N$ gives rise to rapid oscillations with a diagonal matrix of force constants $K > 0 \in \mathbb{R}^{L\times L}$, and $\varepsilon > 0$ is a stiffness parameter. The associated Hamiltonian equations of motion are then given by

$$\begin{aligned} \dot{q} &= p \\ \dot{p} &= -\varepsilon^{-2}G(q)^{\mathrm{T}}Kg(q) - \nabla V(q), \end{aligned} \tag{2}$$

where $G(q) := Dg(q) \in \mathbb{R}^{L\times N}$ denotes the Jacobian matrix of g at q. These equations pose challenges in their numerical treatment as well as for sequential data assimilation techniques in the limit $\varepsilon \to 0$. We observe that solutions of (2) preserve the Hamiltonian energy functional (1) and bounded energy, i.e., $H^\varepsilon(q,p) = \mathcal{O}(1)$ as $\varepsilon \to 0$,

implies $g(q) = \mathcal{O}(\varepsilon)$. In other words, for $\varepsilon \ll 1$ solutions q of bounded energy have to stay close to the constraint manifold

$$\mathcal{M} = \{q \in \mathbb{R}^N : \|g(q)\| = 0\}. \tag{3}$$

From here on we will assume that G is free of critical points so that an explicit local decomposition of q into fast and slow modes is possible.

Lemma 1.1. *Let $\Omega \subseteq \mathbb{R}^N$ be open and bounded and let $\operatorname{rank} G(q) = L$ for all $q \in \Omega$, then the linear map $\mathcal{P}_q : \mathbb{R}^N \to \mathbb{R}^N$ given by*

$$\mathcal{P}_q \coloneqq G^{\mathrm{T}}(q)(G(q)G^{\mathrm{T}}(q))^{-1}G(q) \tag{4}$$

is an orthogonal projection.

Remark 1.2. $\mathcal{P}_q^{\perp}$ denotes the orthogonal projection onto the orthogonal complement of the image of $\mathcal{P}_q$. For every $q \in \mathcal{M}$ its image is included in the corresponding tangent space to $\mathcal{M}$ i.e. $\mathcal{P}_q^{\perp}\mathbb{R}^N \to T_q\mathcal{M}$.

The rapid oscillations orthogonal to the manifold $\mathcal{M}$ are locally characterized by the oscillatory Hamiltonian

$$H^{\varepsilon}_{\mathrm{osc}}(q,p) = \frac{1}{2}p^{\mathrm{T}}\mathcal{P}_q p + \frac{1}{2\varepsilon^2}g(q)^{\mathrm{T}}Kg(q). \tag{5}$$

Remark 1.3. In general the energy (5) is not invariant under the evolution governed by (1).

Henceforth we will assume g to be locally smooth and $G(q)$ to have full rank L for all $q \in \mathbb{R}^N$ satisfying $\|g(q)\| \leq C$ for sufficiently large constant $C > 0$. To state the setting more rigorously, we consider solutions $q^{\varepsilon}, p^{\varepsilon} \in C^1([0,T],\mathbb{R}^N)$ to (2) given the initial conditions

$$\begin{aligned} q^{\varepsilon}(0) &= q_0^0 + \varepsilon\bar{q}, \quad \bar{q} \in \mathbb{R}^N \\ p^{\varepsilon}(0) &= p_0^0 + \varepsilon\bar{p}, \quad \bar{p} \in \mathbb{R}^N. \end{aligned} \tag{6}$$

where $(p_0^0, q_0^0) \in \mathcal{TM}$. Hereby $\mathcal{TM}$ denotes the tangential bundle of $\mathcal{M}$ which we will interpret as manifold in phase space, i.e.,

$$\mathcal{TM} = \{(q,p) \in \mathbb{R}^{2N} : q \in \mathcal{M} \wedge p \in T_q\mathcal{M}\}. \tag{7}$$

Therefore the initial data is, up to a perturbation of order ε, *tangential* [40], and we note in passing that the oscillatory energy (5) is small of order $\mathcal{O}(\varepsilon)$ in this case . In the case of codimension $N - L = 1$ the work of [40, 9]

proved existence of a unique solution q^0, p^0 to the differential algebraic equation system

$$\begin{aligned} \dot{q}^0 &= p^0 & q^0(0) &= q_0^0 \in \mathcal{M} \\ \dot{p}^0 &= -G^{\mathrm{T}}(q^0)K\lambda - \nabla V(q^0) & p^0(0) &= p_0^0 \in T_{q_0^0}\mathcal{M} \\ 0 &= g(q^0) \end{aligned} \tag{8}$$

and convergence $q^\varepsilon(t) \overset{\epsilon\to 0}{\longrightarrow} q^0(t)$, $p^\varepsilon(t) \overset{\epsilon\to 0}{\longrightarrow} p^0(t)$ uniformly for $t \in [0,T]$. The Lagrange multiplier $\lambda \in C^1([0,T], \mathbb{R}^L)$ can be algebraically determined for every $t \in [0,T]$ by

$$0 = \ddot{g}(q^0) = -G(q^0)\left[\nabla V(q^0) + G^{\mathrm{T}}(q^0)K\lambda\right] + \sum_{i=1}^{L}\sum_{j=1}^{L}\left((p^0)^{\mathrm{T}} e_{i,j} p^0\right)\frac{\partial^2 g(q^0)}{\partial q_i \partial q_j}, \tag{9}$$

where e_i is the i-th cartesian unit vector of $\mathbb{R}^N$ and $e_{i,j} = e_i \otimes e_j \in \mathbb{R}^{N\times N}$. Therefore this differential algebraic system is of index 3 and we obtain the additional hidden constraint $G(q^0)p^0 = 0$ by differentiation of $g(q^0) = 0$ with regard to the parameter. Under additional assumptions [5] could prove that solutions $(q^\varepsilon, p^\varepsilon)$, initially ε-close to (q^0, p^0), stay ε-close for exponentially long times.

Example 1.4. Consider a chain of L mass points with positions $r_i \in \mathbb{R}^D$ and momenta $v_i \in \mathbb{R}^D$. The first point (denoted by subscript 0) is assumed to be fixed, all points have equal mass and are under the influence of a constant unidirectional force characterized by a_0. All points are pairwise connected by L (linear) elastic bonds, characterized by their force coefficients $K = \mathrm{diag}(k_1, \dots k_L)$ and their equilibrium lengths $l_i > 0$ for $i = 1 \dots L$. Observing $N = DL$, we can describe the evolution of this mechanical system by (2), if we collect all components of all positions and momenta in q and p, respectively. By means of classical mechanics we then conclude

$$\begin{aligned} g(q) &= \begin{pmatrix} \|r_1\| - l_1 & \dots & \|r_i - r_{i-1}\| - l_i & \cdots & \|r_L - r_{L-1}\| - l_L \end{pmatrix}^{\mathrm{T}} \\ V(q) &= a_0 \sum_{i=1}^{L} e_D^{\mathrm{T}} r_i, \end{aligned} \tag{10}$$

where in this case $e_D \in \mathbb{R}^D$ is the unit vector in the last component. If not stated differently, we assume $D = 2$. Note that this is a genuinely nonlinear model with at least two time scales even for the simplest case of the elastic pendulum with $L = 1$. Due to the work of [4], we expect solutions q^ϵ for this case to stay close to the solutions of the classical pendulum q^0 for long times. Since our goal is to observe and predict chaotic slow dynamics, we wish for a system with solutions q^0, which already exhibit chaotic behavior. For this reason we have to consider a slightly more complex model and choose $L = 2$.

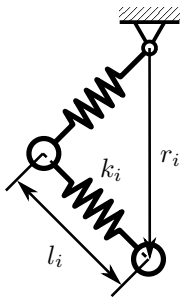

Figure 1: A graphical depiction of (2) using (10) for the case of two mass points $L = 2$ moving in the plane $D = 2$.

As it is of codimension $N - L = 2$ the results of [40] are not longer valid. Nevertheless one can, under additional assumptions, conclude convergence of the solution $\lim_{\varepsilon\to 0} q^\epsilon = q^0$ by the means of [42, 9]. For $L > 1$ and non resonant configurations [5] proved solutions q^ε to stay close to q^0 for long times, if initially so.

Remark 1.5. As demonstrated in [37]

$$g(q^\varepsilon) = \varepsilon^2 \lambda(q^\varepsilon, p^\varepsilon), \tag{11}$$

where $\lambda(q, p)$ is determined by (9), is a better approximation to the slow dynamics of (2) than the zeroth order balance relation $(q, p) \in T\mathcal{M}$. Replacing the constraint in (8) by (11) leads to the concept of soft or flexible constraints as introduced in [37, 45].

Remark 1.6. It should be noted that initial conditions with unconstrained momentum of the form

$$\begin{aligned} q_0^\varepsilon &= q_0^0 + \varepsilon \tilde{q}_0 \qquad & q_0^0 \in \mathcal{M}, \tilde{q}_0 \in \mathbb{R}^N \\ p_0^\varepsilon &= \tilde{p}_0 & \tilde{p}_0 \in \mathbb{R}^N . \end{aligned} \tag{12}$$

lead to an oscillatory energy (5) of order $\mathcal{O}(1)$ instead of $\mathcal{O}(\varepsilon)$. In this case, an additional force term can appear in the limiting equations (8). See [40, 42, 9] for more details.

1.2 Bayesian data assimilation

When describing physical processes by models there are several sources for uncertainties, such as model errors or an uncertainty about the initial conditions. Data assimilation combines model outputs with error prone observational data to estimate the probability distribution of the model state conditioned on the observations. There are two main approaches. First, variational data assimilation estimates a trajectory of the process over an entire time interval by solving a related minimization problem. A well known candidate of this kind is 4D-VAR, as explained e.g. in [39, p. 186].

Another approach is sequential Bayesian data assimilation. This method alternates between a forecast step, in which the probability distribution is evolved in time according to the model until a new observation y_{obs} becomes available. The resulting distribution is called the prior or forecast distribution. The second part of this approach includes an assimilation step which – by applying Bayes' theorem – takes the observations into account and, thus, provides the posterior (or analysis) distribution by

$$\pi^{\text{a}}(z|y_{\text{obs}}) \propto \pi(y_{\text{obs}}|z)\pi^{\text{f}}(z), \tag{13}$$

where here and in the following $z = (q^{\text{T}}, p^{\text{T}})^{\text{T}} \in \mathbb{R}^{2N}$ denotes the state of the model. We denote the posterior and prior density by π^{a} and π^{f} respectively.

In this work we focus on the second approach, and how to apply sequential Bayesian data assimilation to models of the form (2). To this end we consider the deterministic evolution under these model equations, given the normally distributed initial data $z(0) \sim \mathcal{N}(z_0, Q)$ where $z_0 \in T\mathcal{M}$. Furthermore we assume linear observations

$$y_{\text{obs}}(t_k) = H_{\text{obs}} z(t_k) + \zeta \tag{14}$$

where $H_{\text{obs}} \in \mathbb{R}^{I\times 2N}$ is the matrix representing the linear observation map, and $\zeta \sim \mathcal{N}(0, R)$ is the measurement error with Gaussian statistics. Hereby $I \in \mathbb{N}$ is the dimension of the observation space.

For linear models, Gaussian measurement error and Gaussian initial data, the *Kalman filter* solves the problem of matching forecast distribution to the observations optimally [24]. Since the Gaussian structure is exactly preserved in this case, the prior and posterior densities are completely characterized by their means $\bar{z}^{\text{f}}(t_k), \bar{z}^{\text{a}}(t_k)$ and covariances $P^f(t_k), P^a(t_k)$ at time t_k.

When the model equations are nonlinear and therefore the forecast distribution is not Gaussian anymore, we still can recover the main idea of the Kalman filter and approximate $\bar{z}^{\text{f}}(t_k)$ and $P^{\text{f}}(t_k)$ by their empirical counterpart and use the *ensemble Kalman filter* (EnKF) [15] to obtain the posterior mean $\bar{z}^a(t_k)$ and covariance $P^a(t_k)$. To be more specific we draw samples $(z_i(t_{k-1}))_{i\in 1\dots M}$ from $\pi^a(z, t_{k-1})$ and evolve them according to the model equations in time until t_k. Now the resulting $(z_i(t_k))_{i\in 1\dots M}$ are samples of the prior density $\pi^f(z, t_k)$ and we use $\bar{z}^{\text{f}}(t_k) \approx \frac{1}{M}\sum_{i=1}^M z_i(t_k) =: \bar{z}^{\text{f}}$ and $P^{\text{f}}(t_k) \approx \frac{1}{M-1}\sum_{i=1}^M (z_i(t_k) - \bar{z}^{\text{f}})(z_i(t_k) - \bar{z}^{\text{f}})^{\text{T}}$ to estimate the first and second moments of $\pi^f(z, t_k)$. To finally transform the prior samples to samples of the posterior, we assume a linear

transformation (15), but still are left with a choice of the transformation matrix coefficients of σ_{ij} [39]

$$z_j^{\mathrm{a}}(t_k) = \sum_{i=1}^{M} z_i^{\mathrm{f}}(t_k)\sigma_{ij}(t_k) \qquad j = 1, \ldots, M. \tag{15}$$

In the following our choice will be the ensemble square root filter (ESRF) as described e.g. in [39, p.211-212]. The corresponding transfer matrix coefficients are given by

$$\sigma_{ij} = w_i - \frac{1}{M} + S_{ij}, \tag{16}$$

where S and the weights w by

$$S = \left(I + \frac{1}{M-1}(H_{obs}A)^{\mathrm{T}} R^{-1} H_{obs} A \right)^{-\frac{1}{2}} \in \mathbb{R}^{M\times M}, \tag{17}$$

$$w = \frac{1}{M}\sum_{i=1}^{M} e_i - \frac{1}{M-1} S^2 A^{\mathrm{T}} H_{obs}^{\mathrm{T}} R^{-1}(H_{obs}\bar{z}^{\mathrm{f}} - y_{\mathrm{obs}}) \in \mathbb{R}^{M}. \tag{18}$$

Here the ensemble anomalies are denoted by A, i.e.

$$A = \left[z_1^{\mathrm{f}}(t_k) - \bar{z}^{\mathrm{f}}(t_k) \quad \ldots \quad z_M^{\mathrm{f}}(t_k) - \bar{z}^{\mathrm{f}}(t_k) \right] \in \mathbb{R}^{N\times M}. \tag{19}$$

Using a ensemble square root filter, we avoid the perturbation of the observations as necessary for non deterministic versions of the EnKF [44]. Nevertheless our statements do not depend on the specific choice made here.

1.2.1 Failure of the plain ensemble square root filter

Although the Hamiltonian (1) is conserved under the model dynamics (2), i.e.,

$$H^{\varepsilon}(z_i^{\varepsilon\,\mathrm{f}}(t_{k+1})) = H^{\varepsilon}(z_i^{\varepsilon\,\mathrm{a}}(t_k)), \tag{20}$$

it is not conserved under transformation (15) which implements the data assimilation step. In particular, one often observes a severe increase in the oscillatory energy (5), i.e.

$$H^{\varepsilon}_{\mathrm{osc}}(z_i^{\varepsilon\,\mathrm{a}}(t_k)) \gg H^{\varepsilon}_{\mathrm{osc}}(z_i^{\varepsilon\,\mathrm{f}}(t_k)), \tag{21}$$

which, in practice, can lead to a destabilization of the simulation after a few data assimilation cycles. Nevertheless we can control the situation for linear scalar balance relations.

Lemma 1.7. *Let $\sigma \in \mathbb{R}^{M\times M}$ be the transformation matrix of a linear ensemble transform filter. Let $g : \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ be a non negative linear functional. Then for every ensemble of prior samples $q_j^{\mathrm{f}} \in \mathbb{R}^N$ and posterior samples $q_j^{\mathrm{a}} \in \mathbb{R}^N$ with $j \in \{1 \dots M\}$*

$$g\left(q_j^{\mathrm{a}}(t_k)\right) \leq C \max_{i=1\dots m} g\left(q_i^{\mathrm{f}}(t_k)\right) \tag{22}$$

at every time point t_k with $C = \sum_{i=1}^m |\sigma_{ij}|$.

Proof. Due to the linearity of g we can immediately conclude

$$g\left(q_j^{\mathrm{a}}\right) = g\left(\sum_{i=1}^m q_i^{\mathrm{f}}\sigma_{ij}\right) = \sum_{i=1}^m g\left(q_i^{\mathrm{f}}\sigma_{ij}\right) = \sum_{i=1}^m \sigma_{ij} g\left(q_i^{\mathrm{f}}\right) \leq \sum_{i=1}^m |\sigma_{ij}| \max_{l=1\dots m} g\left(q_l^{\mathrm{f}}\right). \tag{23}$$

□

Corollary 1.8. *If the ensemble of prior samples is exactly balanced, i.e. satisfies $g(q_i^{\mathrm{f}}) = 0$ for every $i = 1 \dots M$ then the ensemble of posterior samples will satisfy $g(q_i^{\mathrm{f}}) = 0$ for every $i = 1 \dots M$, too.*

For genuinely nonlinear g we have to give up hope for such results, since we cannot even expect (22) or Corollary 1.8 to hold any longer.

Although not of immediate importance for the current assimilation cycle, the assimilation reduces the mean distance of the ensemble to the observations as expected, but the subsequent forecast can be drastically wrong. In the case of rather small ε this can ultimately lead to filter divergence. An example of this situation is illustrated in Figure 2.

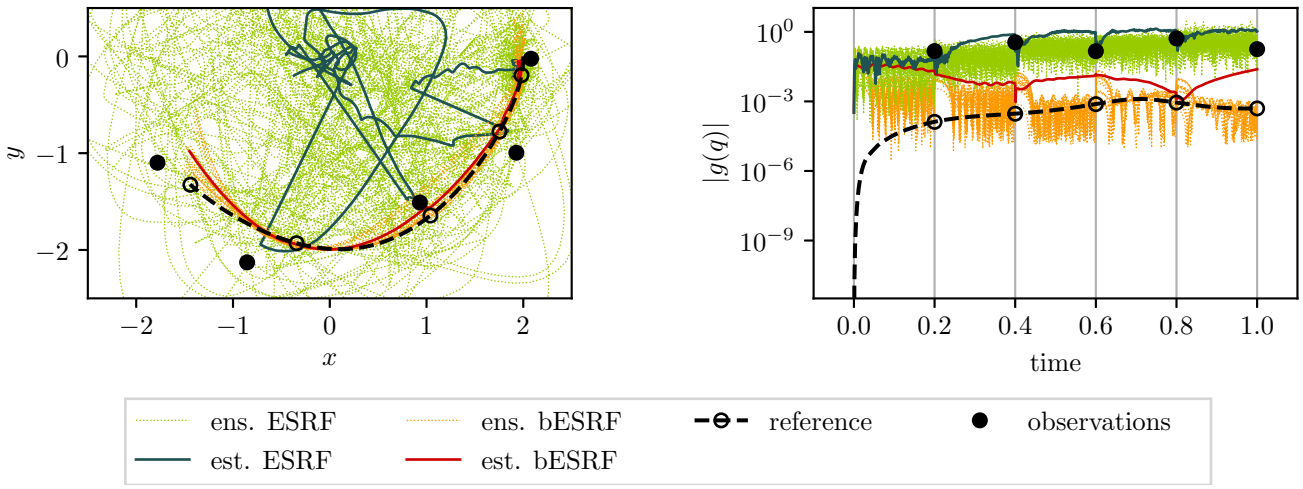

Figure 2: The divergence of an ensemble square root filter (ESRF) applied to the elastic stiff double pendulum as described in Example 1.4. The initial state was chosen as $q_0 = (1, 0, 2, 0)^{\mathrm{T}}$, $p_0 = (0, 0, 0, 0)^{\mathrm{T}}$ and the model parameters according to Table 1. The left figure depicts the ensemble trajectories (ens.) of the second mass point and the associated ensemble means (est.). The plain ESRF diverges after a couple of assimilation cycles, whereas the balanced version (bESRF) performs qualitatively well. The right figure shows the residual $\|g(q)\|$. The residual of the reference solution is non-zero as expected, but stays small. The residual of the balanced ESRF is drastically lower than that of the plain ESRF. We obtain these results using the method proposed in (88) and the setup as described in Chapter 3, c.f. Table 1. In contrast to the situation there, we increase the initial spread and choose $\rho_0 = 0.1$.

2 Proposed methods

To overcome the abovementioned issue, we propose two different methods. The first, subsequently called "penalty method", observes and corrects the balance residuals after the assimilation algorithm. For this purpose we solve a minimization problem structurally similar to the 3DVar method (see e.g. [25]). The second, subsequently called "blended time stepping method", is an extension of ideas first formulated in [3]. This approach does not modify the assimilated states but leverages an intermediate model as part of the forecast step that drives the evolution towards balanced states.

2.1 Ensemble based penalty method

Henceforth we will denote the members of an ensemble $u \in \mathbb{R}^{N\times M}$ by $u_i \in \mathbb{R}^N$ and we use $\bar{u} := \frac{1}{M}\sum_{i=0}^{M} u_i$ for the ensemble mean for notational convenience.

Using this notation, let $\hat{q}_i$ and $\hat{p}_i$ denote the coordinates and momenta of an analysis ensemble provided by

any linear ensemble transform filter. We propose to subsequently apply the current transformation (15) for the ensemble to the forecasted values of g, i.e.

$$\hat{g}_j \coloneqq \sum_i^M g(q_i^{\mathrm{f}})\,\sigma_{ij} \quad j \in \{1,\dots,M\} \tag{24}$$

and minimize the functional $L : \mathbb{R}^{N\times M} \to \mathbb{R}$ after each assimilation procedure as post processing step, where

$$L(u) = \frac{1}{2}\sum_{i=1}^{M}(u_i - \hat{q}_i)^{\mathrm{T}} B(u_i - \hat{q}_i) + S_i(u_i(s)) \tag{25}$$

$$S_i(u_i) = (g(u_i) - \gamma\hat{g}_i)^{\mathrm{T}}\Lambda(g(u_i) - \gamma\hat{g}_i). \tag{26}$$

Hereby $B \in \mathbb{R}^{N\times N}$ is a symmetric positive definite matrix and $\Lambda \in \mathbb{R}^{L\times L}$ is a positive definite diagonal matrix which weights between the importance of the proposed analysis ensemble and the balanced state. Furthermore we choose by $\gamma \in [0,1]$, the amount of balance we wish to achieve relative to the assimilated balance residual $\hat{g}$. The post processed balanced posterior samples then are given as some minimizers, i.e.

$$z^a = \left(\arg\min_u L(u), \hat{p}\right). \tag{27}$$

Remark 2.1. We deliberately refrain from combining the assimilation algorithm and the post processing step for the sake of transparency and to advertise the flexibility of this approach.

2.1.1 Quasi-Newton minimization

The gradient of (25) evaluated at any minimizer thereof vanishes i.e.

$$0 = \frac{\partial L}{\partial u_i}(u) = B(u_i - \hat{q}_i) + \nabla S_i(u_i) \qquad \forall i \in \{1,\dots,M\}. \tag{28}$$

For approximation of this system of coupled nonlinear equations we first consider the Gauss-Newton algorithm, where we first linearize the cost functional (25) by

$$g(u_i) \approx g(u_i^*) + G(u_i^*)(u_i - u_i^*) \tag{29}$$

for some ensemble u^* close to u and therefore obtain

$$\left(B + G^{\mathrm{T}}(u_i^*)\Lambda G(u_i^*)\right)(u_i - u_i^*) \approx -B(u_i^* - \hat{q}_i) - \nabla S_i(u_i^*) \qquad \forall i \in \{1, \dots M\} \tag{30}$$

as condition for critical points. Solving this system of linear equations for the increment ensemble $\Delta u^k \coloneqq u - u^*$ is equivalent to the inversion of $\tilde{B}_i \in \mathbb{R}^{N\times N}$ for $i \in \{1, \dots, M\}$, given by

$$\tilde{B}_i \coloneqq \left(B + G(u_i^*)^{\mathrm{T}}\Lambda G(u_i^*)\right). \tag{31}$$

Each of these matrices is indeed invertible, since B is symmetric positive definite.

This allows us to formulate the update increment for each ensemble member of the increment independently.

$$\Delta u_i^* = -\tilde{B}_i^{-1}\left(B(u_i^* - \hat{q}_i) + \nabla S_i(u_i^*)\right). \tag{32}$$

Remark 2.2. For $L \ll N$ we can reduce the computational costs for each iteration step by expressing $\tilde{B}_i^{-1} = (B + G(u_i^*)^{\mathrm{T}}\Lambda G(u_i^*))^{-1}$ by the Sherman-Morrison-Woodbury formula [20, p. 51].

Subsequently we follow the Quasi-Newton method again and descend along the gradient using

$$u_i^{k+1} - u_i^k = h\Delta u_i^k. \tag{33}$$

Hereby $h > 0$ is a step size determined by a line search for a local minimum of the balance functional L, for details see e.g. [36].

2.1.2 Continuous formulation

Instead of using the gradient descend proposed by the Gauss-Newton methods, we can aim to solve (28) using a different direction of descent. To this end we will introduce a slightly modified search direction, obtained by replacing Λ in (31) by Λh where $h > 0$. We follow (32) and denote the increment obtained by this method by $\Delta\tilde{u}^*$. It is important to observe, that all fixed points of $u^k \mapsto u^k + h\Delta\tilde{u}^k$ are solutions of (28) and therefore again candidates for minimizers. If we drop all the terms of $O(h^2)$ we obtain a stable numerical integration method for a system of differential equations given by

$$\frac{d}{ds}u_i = -(u_i - \hat{q}_i) - B^{-1}G^{\mathrm{T}}(u_i)\Lambda(g(u_i) - \gamma\hat{g}_i) \qquad \forall i \in \{1, \dots, M\} \tag{34}$$

which is then applied to the initial value problem determined by our choice of the initial guess, $u_i = \hat{q}_i$.

Remark 2.3. The evolution governed by (34) is a gradient flow driven by L and the geometry of $\mathrm{diag}(B^{-1}, \dots, B^{-1}) \in \mathbb{R}^{MN \times MN}$. We therefore expect the solution of (34) to converge to an equilibrium solution $u_\infty = \lim_{t\to\infty} u(t)$ satisfying (28).

Proposition 2.4. *The numerical method governed by* (32)

$$u_i^{n+1} = u_i^n - h\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}\left(B(u_i^n - \hat{q}_i) + \nabla S_i(u_i^n)\right) \tag{35}$$

is consistent with (34). *For $h \in (0,1)$ there exists $\delta > 0$ such that every sequence $(u^n)_{n\in\mathbb{N}}$ determined by* (35) *and starting in the open ball $B_\delta(u^\infty)$ converges to an equilibrium solution u^∞.*

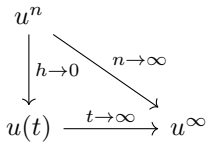

Figure 3: The commuting diagram shows the stability property of discretization (35).

Proof. We recall that $B \in \mathbb{R}^{N\times N}$ is invertible and bounded as it is finite dimensional. The expansion given by the Neumann series now gives

$$\begin{aligned}\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1} &= \big(\mathbb{1} + hB^{-1}G^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}B^{-1} = \sum_{k=0}^{\infty}\big(-B^{-1}G^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^k B^{-1} \\ &= B^{-1} + \mathcal{O}(h).\end{aligned} \tag{36}$$

This allows to conclude consistency by a standard Taylor argument, expanding the solution u at t^n and assuming $u^n = u(t^n)$. For this purpose first rewrite

$$u_i^{n+1} = u_i^n - h\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}\frac{\partial L}{\partial u_i}(u^n) \tag{37}$$

$$= u_i^n - hB^{-1}\frac{\partial L}{\partial u_i}(u^n) + \mathcal{O}(h^2) \tag{38}$$

and subsequently conclude

$$\|u_i(t^{n+1}) - u_i^{n+1}\| = \|u_i(t^n) - hB^{-1}\frac{\partial L}{\partial u_i}(u(t^n)) - u_i^n + hB^{-1}\frac{\partial L}{\partial u_i}(u^n) + \mathcal{O}(h^2)\| = \|O(h^2)\|. \tag{39}$$

This implies global first order consistency and the first part of the statement. For the second part let $h \in (0,1)$. Subtracting by u_i^∞ on both sides and furthermore using

$$0 = B(u_i^\infty - \hat{q}_i) + G(u_i^n)\Lambda(g(u_i^n) - \bar{\hat{g}}) \tag{40}$$

allows us to conclude equivalence of (35) and the following identity.

$$\begin{aligned} u_i^{n+1} - u_i^\infty &= u_i^n - u_i^\infty - h\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}\left(B(u_i^n - u_i^\infty) + G^{\mathrm{T}}(u_i^n)\Lambda(g(u_i^n) - g(u_i^\infty))\right) \\ &= \Big(\mathbb{1} - h\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}\left(B + G^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\right)\Big)(u_i^n - u_i^\infty) + h\, r(u_i^n, u_i^n - u_i^\infty) \end{aligned} \tag{41}$$

The last equality is valid as long as $u^n \in B_\rho(u^\infty)$ for sufficiently small $\rho > 0$. In this case we can apply Taylor expansion which also gives us

$$R(w_i) \coloneqq \sup_{v \in B_\rho(u^\infty)} \|r(v_i, w_i)\| \in \mathcal{O}(\|w_i\|^2). \tag{42}$$

The fact that B and $G(u_i^n)^{\mathrm{T}}\Lambda G(u_i^n)$ are both symmetric positive definite allows us to conclude the following estimate

$$\begin{aligned} \|u_i^{n+1} - u_i^\infty\| &\le \|\,\mathbb{1} - h\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}\left(B + G^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\right)\|\|u_i^n - u_i^\infty\| + hR(u_i^n - u_i^\infty) && (43)\\ &\le (1-h)\|\big(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n)\big)^{-1}B\|\|u_i^n - u_i^\infty\| + hR(u_i^n - u_i^\infty) && (44)\\ &\le \frac{(1-h)\|B\|}{\|(B + hG^{\mathrm{T}}(u_i^n)\Lambda G(u_i^n))\|}\|u_i^n - u_i^\infty\| + hR(u_i^n - u_i^\infty) && (45)\\ &\le (1-h)\|u_i^n - u_i^\infty\| + hC_1\|u_i^n - u_i^\infty\|^2 && (46)\\ &\le (1 - h + hC_1\|u_i^n - u_i^\infty\|)\|u_i^n - u_i^\infty\| && (47)\\ &\le C_2\|u_i^n - u_i^\infty\|. && (48) \end{aligned}$$

Hereby the constant satisfies $C_1 < 1$ as long as u_i^n is already close enough to u_i^∞. If so, then we immediately obtain $\left\|u_i^{n+1} - u_i^\infty\right\| C_1 < 1$ and therefore $C_2 < 1$ too. An inductive argument finally implies convergence to the equilibrium solution u_i^∞ for sufficiently close initial value. □

2.1.3 Gradient free approximation

Assume g to be the observation operator and let $\overline{\hat{g}}$ be a constant observation, then the time evolution of the Ensemble-Kalman-Bucy filter, introduced in [6], is determined by

$$\frac{d}{ds}u_i = -P_{uu}^{-1}(G^{\mathrm{T}}(\bar{u})(g(\bar{u}) - \overline{\hat{g}}) + G^{\mathrm{T}}(u_i)(g(u_i) - \overline{\hat{g}})) \qquad \forall i \in \{1, \dots, M\}. \tag{49}$$

Hereby P_{uu} denotes the empirical covariance matrix

$$P_{uu} \coloneqq \frac{1}{M-1} \sum_{i=0}^{M} (u_i - \overline{u}) (u_i - \overline{u})^{\mathrm{T}} . \tag{50}$$

If we consider B in (34) to be the inverse background error covariance matrix we can approximate $B^{-1} \approx P_{uu}$ to obtain evolution equations which share some structure with the Ensemble-Kalman-Bucy filter with nonlinear observation operator g. Motivated by this analogy to the Ensemble-Kalman-Bucy filter we argue that it is reasonable and computationally more efficient to use a modified ensemble approximation of the Kalman-Bucy filter in the spirit of (35), as we are more interested in the reduction of the residual of g for each ensemble member than in finding an accurate solution to (25). As e.g. pointed out in [23] we can compute such an ensemble approximation observation matrix free i.e. without the evaluation of the gradient G at the expense of certain linearizations. For this purpose we reconsider Remark 2.2 and express the inverse by

$$\tilde{B}_i^{-1} = (B + G^{\mathrm{T}}(u_i^*)\Lambda G(u_i^*))^{-1} = B^{-1} - B^{-1}G^{\mathrm{T}}(u_i^*)\Lambda^{1/2}(\mathbb{1} + \Lambda^{1/2}G(u_i^*)B^{-1}G^{\mathrm{T}}(u_i^*)\Lambda^{1/2})^{-1}\Lambda^{1/2}G(u_i^*)B^{-1}. \tag{51}$$

We linearize and introduce an error of the same order of magnitude as the ensemble spread by the use of the following empirical estimates (c.f. [14])

$$B^{-1} \approx P_{uu} \coloneqq \frac{1}{M-1} \sum_{i=0}^{M} (u_i - \overline{u}) (u_i - \overline{u})^{\mathrm{T}} , \tag{52}$$

$$B^{-1}G^{\mathrm{T}}(\overline{u}^*) \approx B^{-1}G^{\mathrm{T}}(u_i^*) \approx P_{ug} \coloneqq \frac{1}{M-1} \sum_{i=0}^{M} (u_i - \overline{u}) \left(g(u_i) - \overline{g(u_j)}\right)^{\mathrm{T}} , \tag{53}$$

$$G(u_i^*)B^{-1}G^{\mathrm{T}}(u_i^*) \approx G(u_i^*)B^{-1}G^{\mathrm{T}}(\overline{u}^*) \approx P_{gg} \coloneqq \frac{1}{M-1} \sum_{i=0}^{M} \left(g(u_i) - \overline{g(u_j)}\right) \left(g(u_i) - \overline{g(u_j)}\right)^{\mathrm{T}} . \tag{54}$$

These approximations drastically simplify (51) which now reads

$$\tilde{B}_i^{-1} \approx P_{uu} - P_{ug}\Lambda^{1/2}(\mathbb{1} + \Lambda^{1/2}P_{gg}\Lambda^{1/2})^{-1}\Lambda^{1/2}P_{ug}^{\mathrm{T}}. \tag{55}$$

As we substitute into (32), we finally obtain a gradient free version of the increment

$$\begin{aligned} \Delta u_i^k \approx &- \left(P_{uu} - P_{ug}\Lambda^{1/2}(\mathbb{1} + \Lambda^{1/2}P_{gg}\Lambda^{1/2})^{-1}\Lambda^{1/2}P_{ug}^{\mathrm{T}}\right) P_{uu}(u_i^* - \hat{u}_i) \\ &- \left(P_{ug} - P_{ug}\Lambda^{1/2}(\mathbb{1} + \Lambda^{1/2}P_{gg}\Lambda^{1/2})^{-1}\Lambda^{1/2}P_{gg}\right) \Lambda \left(g(u_i^*) - \hat{g}_i\right) . \end{aligned} \tag{56}$$

2.2 Blended time-stepping

Motivated by the results of [3], we introduce a numerical time stepping scheme that extends a classical projection approach by subsequent blending steps. These steps continuously blend between the reduced limit model (8) and the unconstrained model (2). This approach was originally developed in the context of incompressible fluid dynamics where the singular perturbation arises by the vanishing Mach number limit $\mathrm{Ma} \to 0$. In addition to the classical projection schemes introduced by [12], much effort was spent on developing asymptotic preserving low Mach number numerical schemes and variants thereof. The essential point is their ability to blend between the (weakly) compressible and the incompressible dynamics without additional stability constraints. As observed in [3], solving the incompressible model immediately after the assimilation and subsequently blending smoothly and over a few time steps back to the compressible one, can further reduce artificial imbalances caused by data assimilation, relative to an approach that simply projects the system state onto the incompressible manifold in one step and then proceeds with the compressible model.

To adapt this strategy to our situation, we would like to investigate a numerical method

$$z^{n+1} = \psi_h^\alpha(z^n) \qquad \alpha \in [0,1]. \tag{57}$$

For the case $\alpha = 0$ we aim to obtain a projection method, which keeps the momenta tangential and is consistent with the constrained system (8). For the case $\alpha = 1$ the method should resolve the unconstrained model (2). In between, i.e., for $\alpha \in (0,1)$, we suggest to follow a dissipative model to be introduced below. In this approach we accept a non vanishing consistency error with respect to the fast model when evolving the system with $\alpha \in [0,1)$. As discussed in the beginning, however, we can assume the solutions of the unconstrained system to stay ε – close to the solutions of the constrained one. This enables us to locally decompose the consistency error into two parts, one in $\mathcal{M}$, caused by the nonlinearity of V and another orthogonal to $\mathcal{M}$. The slow first part is assumed to be captured by the data assimilation, whereas the second fast part has, as discussed before, only magnitude of order $\mathcal{O}(\varepsilon)$. Given artificial imbalances, the latter is not necessarily true. For this purpose the discrete evolution of the blended method will dampen the fast oscillations orthogonal to $\mathcal{M}$ rapidly, as long as $\alpha \in (0,1)$ and until they attain the correct amplitude of $\mathcal{O}(\varepsilon)$. To this end we propose to use the blending method (57) as follows. We denote the *blending window* by $k \geq 1$ and start our forecast at time t_n. Let η be the number of forecast time integration steps, then the following two steps are repeated in every forecast cycle (c.f. Figure 4).

1. **Blending:** Let $\alpha \in \mathbb{R}^k$ such that $0 = \alpha_1 < \alpha_2 \leq \cdots \leq \alpha_{k-1} < \alpha_k = 1$. Integrate until t_{n+k} using

$$q_{n+k} = \left(\psi_h^{\alpha_k} \circ \psi_h^{\alpha_{k-1}} \circ \cdots \circ \psi_h^{\alpha_1}\right)(q_n) \tag{58}$$

2. **Forecast:** Obtain forecast at $t_{n+\eta}$ by evolving q_{n+k} along ψ_h^1 for $\eta - k$ – times.

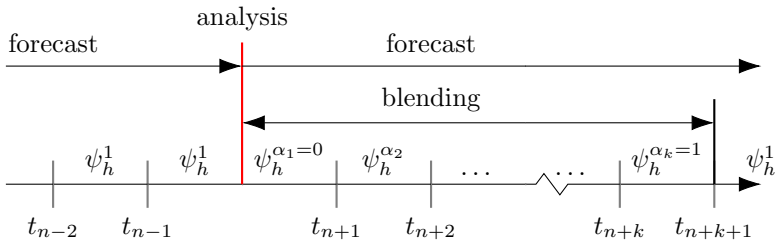

Figure 4: Blended time stepping applied after analysis with blending window k. The numerical flow ψ_h^α with step width h is given by (57).

Figure 5 illustrates the qualitative behaviour of the blended time stepping for the stiff elastic double pendulum, introduced in Example 1.4, with slightly unbalanced initial coordinates. We observe that with respect to the balance residual the blended time stepping improves the situation drastically. After short time the residuals of the initially unbalanced and initially balanced solution match. Since we dissipate energy in the fast variables (c.f. Lemma 2.8) as long as $\alpha \notin \{0, 1\}$, the overall energy of the system decreases as expected and the slow rotational motion of the stiff double pendulum is therefore resolved reasonably well with regard to balance and energy. Nevertheless, due to the lack of a priori knowledge so far, we chose $\alpha \in [0, 1]$ continuously, but as we already can guess from the form of the decay, this is a brute force and suboptimal choice in the sense that we can find a smaller range of α within which the solution relaxes to the slow motion more quickly. We leave the development of an optimized control of the blending sequence for future work.

2.2.1 Model hierarchy

We aim to understand the proposed strategy in terms of a model hierarchy. For this purpose we first introduce and discuss a dissipative version of model (2) given by

$$\begin{aligned} \dot{\tilde{q}} &= \tilde{p} \\ \dot{\tilde{p}} &= -\varepsilon^{-2} G(\tilde{q})^{\mathrm{T}} K g(\tilde{q}) - d\mathcal{P}_{\tilde{q}}\tilde{p} - \nabla V(\tilde{q}), \end{aligned} \tag{59}$$

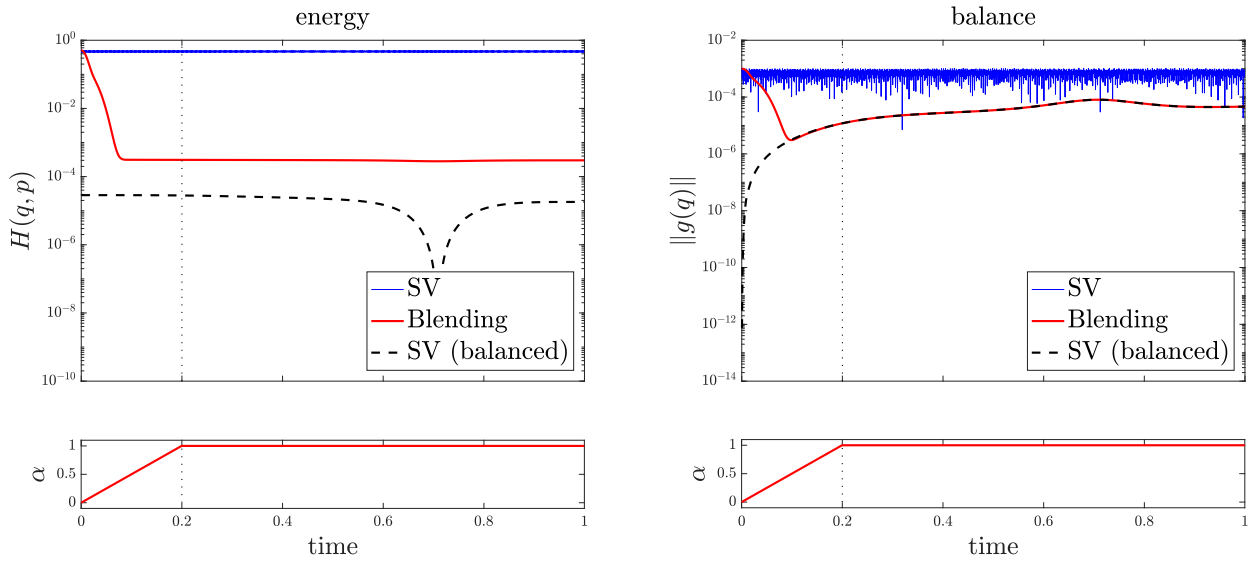

Figure 5: Energy, residual of balance relation and solution for the stiff elastic double pendulum c.f. Example 1.4. Initially unbalanced numerical solutions based on the Störmer-Verlet (SV) (blue) and the blending method (red), respectively. The corresponding parameter α for the blended time stepping is shown and the start of the pure forecasting region is marked by a vertical dotted line. As the reference we display the data for an initially balanced solution, computed again by the Störmer-Verlet method (dashed).

where d is an additional scalar damping coefficient. The direction of the damping is chosen to enable consistency with the original model (2) and was first proposed in the context of stabilization techniques in [18]. Since fast and slow energy parts of the Hamiltonian (1) can be separated only by an asymptotic argument and are coupled nonlinearly, we do not expect to completely dissipate fast energy of (1) by the evolution of (59). Nevertheless in reasonably well separated cases the impact will be negligible, especially in the context of data assimilation, where the correct slow energy itself is known only up to some random perturbation. More concretely we will argue that trajectories produced by a blended method will locally relax to the constraint manifold by similar means as in the context of stabilization of differential algebraic equations [2]. Henceforth we will consider (8) in a slightly more general form.

$$\begin{aligned} \dot{q} &= p & q(0) &= q_0 \in \mathbb{R}^N \\ \dot{p} &= -G(q)^{\mathrm{T}} K\lambda - \nabla V(q) & p(0) &= p_0, \quad G(q_0)p_0 = 0 \\ g(q) &= g(q_0) \end{aligned} \tag{60}$$

Remark 2.5. After differentiating the constraint $G(q)p = 0$, λ is given as before by (9). For $g(q_0) = 0$ this system is equivalent to the constrained system (8) in the sense that $\mathcal{M}$ is invariant under the evolution in time following (60). Due to continuous dependency on initial data we furthermore conclude that solutions to (60) approach solutions

of (8) as $g(q_0) \to 0$.

To develop an intuition regarding the behavior of solutions to (59) we discuss the arguably simplest model in the class of such problems with multiple scales, the uncoupled harmonic oscillator.

Example 2.6 (Damped harmonic oscillator). Let $q = (\mu, \nu)^{\mathrm{T}}$, $p = (\eta, \zeta)^{\mathrm{T}}$, $K = \mathrm{diag}(\varepsilon^2, 1)$ and $V(q) = 0$

$$\dot{\mu} = \eta \qquad \dot{\eta} = -\mu \tag{61}$$

$$\dot{\nu} = \zeta \qquad \dot{\zeta} = -\frac{1}{\varepsilon^2}\nu - 4d\zeta \tag{62}$$

The well known analytical solutions for the damped harmonic oscillator are given by $\mu = \mu_0 \cos(t) + \nu_0 \sin(t)$ and

$$\nu = \begin{cases} e^{-2dt}\left(\nu_0 + t(2d\nu_0 + \zeta_0)\right) & 2d = \frac{1}{\varepsilon} \\ e^{-2dt}\left((\nu_0 + \frac{\zeta_0}{\omega_d})e^{i\omega_d t} + (\nu_0 - \frac{\zeta_0}{\omega_d})e^{-i\omega_d t}\right) & 2d \neq \frac{1}{\varepsilon} \end{cases} \tag{63}$$

where the frequency for the fast damped component is given by $\omega_d = \sqrt{\frac{1}{\varepsilon^2} - 4d^2}$. We immediately realize that $d = 0$ gives us the solution for the highly oscillatory system (2) and furthermore relaxes to the constraint (in this case also slow) manifold exponentially fast. In the general nonlinear and coupled case we present the corresponding result in Lemma 2.8.

For the overdamped limit, i.e., for $d\varepsilon \to \infty$ as $\varepsilon \to 0$ and $d \to \infty$, we conclude uniform convergence to the same solution $(\mu, 0)$ as for the constrained system as long as $\nu_0 \in o(1/d)$ and $\zeta_0 \in o(1)$. Again this result can be stated for in more general form and is presented in Lemma 2.10.

The following lemma summarizes the well known (c.f. [38]) split of variables into a slow tangential and a fast normal part. It will enable us to identify slow and fast variables with respect to the different asymptotic limits.

Lemma 2.7. *Any solution (q, p) of System* (59) *can be split into components μ, η and ν, ζ which satisfy*

$$\begin{aligned} q &= E^{\mathrm{T}}\mu + G^{\mathrm{T}}(GG^{\mathrm{T}})^{-\frac{1}{2}}\nu \\ p &= E^{\mathrm{T}}\eta + G^{\mathrm{T}}(GG^{\mathrm{T}})^{-\frac{1}{2}}\zeta\,. \end{aligned} \tag{64}$$

In the new coordinates ((μ, η), (ν, ζ)), system (59) *is equivalent to*

$$\dot{\mu} = \dot{E}q + \eta \qquad \dot{\eta} = \dot{E}p - E\nabla V(q) \tag{65}$$

$$\dot{\nu} = \left(\frac{d}{dt}(GG^{\mathrm{T}})^{-\frac{1}{2}}G\right)p + \zeta \qquad \dot{\zeta} = \left(\frac{d}{dt}(GG^{\mathrm{T}})^{-\frac{1}{2}}G\right)p - \varepsilon^{-2}(GG^{\mathrm{T}})^{\frac{1}{2}}Kg(q) - d\zeta - (GG^{\mathrm{T}})^{-\frac{1}{2}}G\nabla V(q). \tag{66}$$

Furthermore the fast energy H_{osc} satisfies

$$H^{\varepsilon}_{osc}((\nu,\zeta),(\mu,\eta)) = \frac{1}{2}\zeta^{\mathrm{T}}\zeta + \frac{1}{2\varepsilon^2}g(q(\mu,\nu))^{\mathrm{T}}Kg(q(\mu,\nu)) \tag{67}$$

Proof. For the sake of readability we will omit the argument q for G, E throughout the proof. We split momenta tangential and orthogonal to $\mathcal{M}$ denoted by η and ζ as well as the coordinates denoted by μ and ν, respectively. More concretely we choose

$$\begin{aligned} \mu &= Eq & \eta &= Ep \\ \nu &= (GG^{\mathrm{T}})^{-\frac{1}{2}}Gq & \zeta &= (GG^{\mathrm{T}})^{-\frac{1}{2}}Gp \end{aligned} \tag{68}$$

where the columns of $E \in \mathbb{R}^{(N-L)\times N}$ are an orthonormal basis of $\left(\mathcal{P}_q\mathbb{R}^N\right)^{\perp}$. It is easy to check that $E^{\mathrm{T}}E$ is a orthogonal projection onto $\left(\mathcal{P}_q\mathbb{R}^N\right)^{\perp}$ and since orthogonal projections onto a fixed subspace are unique (for every q), we already know $E^{\mathrm{T}}E = \mathcal{P}_q^{\perp} = \mathbb{1} - \mathcal{P}_q$. Substituting (68) into the right hand side of (64) we get

$$E^{\mathrm{T}}Eq + G^{\mathrm{T}}(GG^{\mathrm{T}})^{-1}Gq = \mathcal{P}_q^{\perp}q + \mathcal{P}_q q = q. \tag{69}$$

Since we used the same geometry to split the momenta this already implies (64).

We differentiate (68) and use system (59) for $\dot{q}$ and $\dot{p}$. Since by construction $G^{\mathrm{T}}E = 0 = E^{\mathrm{T}}G$ most of the terms drop and we conclude (65) and (66) after some straightforward algebraic manipulation. Recalling GG^{T} is symmetric positive definite, the last statement (67) finally follows from

$$H^{\varepsilon}_{osc}(q,p) = \frac{1}{2}p^{\mathrm{T}}\mathcal{P}_q p + \frac{1}{2\varepsilon^2}g(q)^{\mathrm{T}}Kg(q) \tag{70}$$

$$= \frac{1}{2}p^{\mathrm{T}}G^{\mathrm{T}}(GG^{\mathrm{T}})^{-\frac{1}{2}^{\mathrm{T}}}(GG^{\mathrm{T}})^{-\frac{1}{2}}Gp + \frac{1}{2\varepsilon^2}g(q)^{\mathrm{T}}Kg(q) \tag{71}$$

$$= \frac{1}{2}\zeta^{\mathrm{T}}\zeta + \frac{1}{2\varepsilon^2}g(q)^{\mathrm{T}}Kg(q). \tag{72}$$

□

Lemma 2.8. *Let $d = \frac{c}{\varepsilon}$, with $c > 0$ fixed. Then solutions to (59) which initially satisfy $H^{\varepsilon}_{osc}(q,p) \in \mathcal{O}(\varepsilon^{-2})$, dissipate fast energy down to some residual of order $H^{\varepsilon}_{osc} \in \mathcal{O}(1)$, if only ε is sufficiently small.*

Proof. Again we will omit the arguments of E and G for notational convenience. Additionally we introduce $\Gamma \coloneqq (GG^{\mathrm{T}})^{-\frac{1}{2}}G$. We will prove the statement by arguments from geometric singular perturbation theory [16]. For this purpose we split system (59) into slow and fast parts by the means of Lemma 2.7. Subsequently we multiply by ε

and rescale $\hat{\zeta} = \varepsilon\zeta$ which results in

$$\dot{\mu} = \dot{E}q + \eta \qquad \dot{\eta} = \dot{E}p - E\nabla V(q) \tag{73}$$

$$\varepsilon\dot{\nu} = \varepsilon\dot{\Gamma}q + \hat{\zeta} \qquad \dot{\hat{\zeta}} = \varepsilon^2\dot{\Gamma}p - (GG^{\mathrm{T}})^{\frac{1}{2}}Kg(q) - \varepsilon d\hat{\zeta} - \varepsilon^2(GG^{\mathrm{T}})^{-\frac{1}{2}}G\nabla V(q). \tag{74}$$

We denote the right hand side of the fast variables by

$$F((\nu,\hat{\zeta}),(\mu,\eta),\varepsilon(d)) = \begin{pmatrix} \varepsilon\dot{\Gamma} + \hat{\zeta} \\ -\varepsilon^2\dot{\Gamma}p - (GG^{\mathrm{T}})^{\frac{1}{2}}Kg(q) - \varepsilon d\hat{\zeta} - \varepsilon^2(GG^{\mathrm{T}})^{-\frac{1}{2}}G\nabla V(q) \end{pmatrix} \tag{75}$$

In the limit $d\varepsilon = \text{const.}$, $\varepsilon \to 0$ we identify the critical manifold as

$$\mathcal{T}\hat{\mathcal{M}} \coloneqq \left\{ (\eta,\hat{\zeta},\nu,\mu) \in \mathbb{R}^{2N} : g(q(\mu,\nu)) = 0 \wedge \hat{\zeta} = 0 \right\} . \tag{76}$$

Next we prove normal hyperbolicity of the critical manifold, i.e., we show that there are no eigenvalues of $\frac{\partial F}{\partial(\nu,\hat{\zeta})}$ with vanishing real part. The gradient evaluated on the manifold $\mathcal{T}\hat{\mathcal{M}}$ and for $\varepsilon = 0$ is given by the block matrix

$$DF \coloneqq \frac{\partial}{\partial(\nu,\hat{\zeta})}F((\nu,\hat{\zeta}),(\mu,\eta),\varepsilon)|_{(\nu,\hat{\zeta})\in\mathcal{T}\hat{\mathcal{M}},\varepsilon=0} = \begin{pmatrix} 0 & \mathbb{1} \\ -(GG^{\mathrm{T}})^{\frac{1}{2}}K(GG^{\mathrm{T}})^{\frac{1}{2}} & -\mathbb{1} \end{pmatrix}. \tag{77}$$

To compute the eigenvalues of this non symmetric matrix, we first recall that $(GG^{\mathrm{T}})^{\frac{1}{2}}$ is symmetric positive definite and since K is a strictly positive diagonal matrix, $\tilde{K} \coloneqq (GG^{\mathrm{T}})^{\frac{1}{2}}K(GG^{\mathrm{T}})^{\frac{1}{2}}$ is symmetric and positive definite, i.e., it has L positive eigenvalues $\omega_{\tilde{K},j} > 0$. Therefore we conclude zero is no eigenvalue of DF by $\det DF = \det(-\mathbb{1})\det(-\tilde{K})$. Using the Schur complement again we argue for some eigenvalue $\omega \neq 0$ of DF

$$\det(DF - \omega\,\mathbb{1}) = \det(-\omega\,\mathbb{1})\det(-(1+\omega)\,\mathbb{1} - \frac{1}{\omega}\tilde{K}). \tag{78}$$

The determinant vanishes if and only if there is $j \in \{1,\dots,L\}$ such that

$$-\omega(\omega+1) = \omega_{\tilde{K},j}. \tag{79}$$

Solving this quadratic equation already provides us with all possible eigenvalues by

$$\omega_{\pm,j} = \frac{-1 \pm \sqrt{1 - 4w_{\tilde{K},j}}}{2}. \tag{80}$$

We directly observe $\mathrm{Re}\,\omega_{\pm,j} < 0$ for all $j \in \{1, \dots L\}$ and therefore notice that $\mathcal{T}\hat{\mathcal{M}}$ is normally hyperbolic. By finally applying Fenichel's theorem we obtain existence of slow manifolds S^ε (c.f. [27]) ε-close to a compact submanifold $S \subset \mathcal{T}\hat{\mathcal{M}}$ of the critical one as long as ε is sufficiently small. More specifically we conclude for any $((\eta^\varepsilon, \zeta^\varepsilon), (\mu^\varepsilon, \nu^\varepsilon)) \in S^\varepsilon$

$$g(q(\mu^\varepsilon, \nu^\varepsilon))^{\mathrm{T}} K g(q(\mu^\varepsilon, \nu^\varepsilon)) \leq c_1 \varepsilon^2 \tag{81}$$

$$\hat{\zeta}^{\varepsilon\mathrm{T}} \hat{\zeta}^\varepsilon \leq c_2 \varepsilon^2 \tag{82}$$

and therefore

$$\max_{(\eta^\varepsilon,\zeta^\varepsilon),(\mu^\varepsilon,\nu^\varepsilon))\in S} H^\varepsilon_{osc}((\eta^\varepsilon, \zeta^\varepsilon), (\mu^\varepsilon, \nu^\varepsilon)) \in \mathcal{O}(1). \tag{83}$$

Another consequence of Fenichel's theorem is that the dynamical behaviour of the linearization DF of the fast subsystem on the critical manifold already determines the dynamical behaviour of solutions starting off a slow manifold S^ε. Since all eigenvalues of DF have negative real part, we conclude S as well S^ε is attracting. Therefore any solution starting nearby will approach some S^ε which finally implies the energy dissipation as stated. □

Subsequently we will use (59) to establish a model hierarchy which resembles the analytical counterparts discretized by the blended numerical method (57). The following two lemmata concern the behaviour of the limit cases $d \to 0$ and $d \to \infty$. The first one is based on the classical result of continuous dependency on initial data and parameters for ordinary differential equations with continuously differentiable right hand side. In both cases we fix $\varepsilon > 0$ and omit this standard proof.

Lemma 2.9. *Let $\varepsilon > 0$ be fixed. Solutions $(\tilde{q}, \tilde{p})$ of the dissipative system (59) approach solutions of the purely Hamiltonian system (2) as $d \to 0$.*

For the other part $d \to \infty$ we use again geometric singular perturbation theory and we can conclude a slightly different type of statement in terms of invariant manifolds.

Lemma 2.10. *For sufficiently large d and $\varepsilon^2 \in o(1/d)$ there exists a Manifold $\mathcal{M}_{1/d}$ which lies within $\mathcal{O}(1/d)$ of any compact subset of $\mathcal{M}_\infty := \{(q,p) \in \mathbb{R}^{2N} : \zeta(q,p) = 0\}$ (c.f. Lemma 2.7) and is invariant under the evolution of (59). Furthermore every solution starting off, but sufficiently close to this subset will approach $\mathcal{M}_{1/d}$.*

Proof. As pointed out we aim to apply geometric singular perturbation theory again. Therefore we start as before by splitting slow and fast momenta explicitly utilizing Lemma 2.7. Contrary to the situation in Lemma 2.8 the coordinates then are both slow variables. By dividing the momentum equation in (66) by d and passing to the limit $d \to \infty$ we obtain the critical manifold as $\mathcal{M}_\infty = \{(q,p) \in \mathbb{R}^{2N} : \zeta(q,p) = 0\}$. We denote the right hand side of the momentum equation in (66) by $F((\eta,\zeta), q, 1/d)$ and linearize on $\mathcal{M}_\infty$.

$$DF \coloneqq \frac{\partial}{\partial \zeta} F((\eta,\zeta), q, 1/d)|_{(\nu,\hat{\zeta})\in\mathcal{M}_\infty, \varepsilon=0} = -\left(GG^{\mathrm{T}}\right) \tag{84}$$

Since GG^{T} is positive definite and by assumption $\operatorname{rank}(GG^{\mathrm{T}}) = L$ we conclude that $-\left(GG^{\mathrm{T}}\right)^{1/2}$ has exactly L negative Eigenvalues. M_∞ is therefore normally hyperbolic and we now infer by Fenichel's theorem [27] existence of an invariant (with respect to (59)) manifold $\mathcal{M}_{1/d}$, $1/d$ – close to a compact subset of our choice of $\mathcal{M}_\infty$, exactly as stated. Since we additionally have only a stable subspace on $\mathcal{M}_\infty$ we gain the attractive behavior of $\mathcal{M}_{1/d}$ by the same theorem. □

Remark 2.11. The same statement is true if we take a compact submanifold of $\mathcal{M}_\infty$ and therefore also cover the case where the evolution starts on the constraint manifold $\mathcal{M} \subseteq \mathcal{M}_\infty$.

Corollary 2.12. *For sufficiently large d there exists a Manifold $\mathcal{M}_{1/d}$ which lies within $\mathcal{O}(1/d)$ of any compact subset of $\mathcal{M}$ and is invariant under the evolution of* (59)*. Furthermore every solution starting off but sufficiently close to this subset will approach $\mathcal{M}_{1/d}$.*

Remark 2.13. Although the preceding corollary tells us there is a at least one slow manifold for large d that satisfies the constraint, this does not imply we approach one of this kind, when starting slightly off $\mathcal{M}$.

So far we have only considered the analytical properties of the dissipative system (59). Building upon the insights gained, we now propose a related numerical method.

2.2.2 Discretization

As already pointed out previously our method is supposed to be consistent with the unconstrained system (2) and system (8) for $\alpha = 1$ and $\alpha = 0$ respectively. For the first case we furthermore require conservation of the energy H for discrete solutions as well and choose the method to be symplectic i.e. such that the gradient of the discrete flow $D\psi_h^1$ satisfies

$$(D\psi_h^1)^{\mathrm{T}} J (D\psi_h^1) = J \coloneqq \begin{pmatrix} 0 & \mathbb{1} \\ -\mathbb{1} & 0 \end{pmatrix}. \tag{85}$$

This property is shared with the analytical flow and responsible for exact conservation of the energy functional H for analytical solutions as well as preservation of a modified energy functional for discrete solutions. For an extensive presentation and discussion on this topic see e.g. [22] or [28].

We build our method on the symplectic (c.f. (85)) Störmer Verlet method (86). Despite its simplicity this method performs exceptionally well and is extensively discussed in detail in e.g. [22].

$$\begin{aligned} q^{n+\frac{1}{2}} &= q^n + h\, p^n \\ p^{n+1} &= p^n - h\nabla V(q^{n+\frac{1}{2}})) - \frac{h}{\varepsilon^2} G^{\mathrm{T}}(q^{n+\frac{1}{2}}) K g(q^{n+\frac{1}{2}}) \\ q^{n+1} &= q^{n+\frac{1}{2}} + h\, p^{n+1} \end{aligned} \tag{86}$$

In contrast to (2), the constrained model equations (8) are in fact a system of differential algebraic equations of index 3 [1]. Solving these equations numerically leaves the choice to fulfill the constraint exactly or accept a numerical approximation error for $g(q^n) = 0$. In the Hamiltonian context, the first choice suggests the SHAKE and RATTLE schemes (c.f. [28]). Given initial states q^n and tangential momenta p^n, both algorithms use a projection to ensure $q^{n+1} \in \mathcal{M}$ and $p^{n+1} \in T_{q^{n+1}}\mathcal{M}$.

The second approach for solving a system of differential algebraic equations is to accept approximation errors for the constraint itself. This approach relies on index reduction of the analytical system and subsequent discretization c.f. e.g. [1]. In this context a common task is to design stabilized methods [2] which allow for a discrete evolution close to the constraint manifold such that the error on the constraint stays small for long times.

Since we do not aim to run the constrained model $\alpha = 0$ for more than a few time steps in the blended method, we will employ an index reduction approach but ignore the issue of stabilization at this point. Motivated by the Störmer-Verlet method we propose a projection method for (8) which satisfies the hidden constraint

$$G(q)p = 0 \tag{87}$$

up to a given tolerance and the constraint $g(q) = 0$ in (8) up to a global error of order $\mathcal{O}(h^2)$. The proposed blended method reads

$$\begin{aligned} q^{n+\frac{1}{2}} &= q^n + \frac{h}{2} p_n \\ p^{n+1,\alpha} &= p^n - h\nabla V(q^{n+\frac{1}{2}}) - hG^{\mathrm{T}}(q^{n+\frac{1}{2}}) \left(\frac{\alpha}{\varepsilon^2} K g(q^{n+\frac{1}{2}}) + (1-\alpha)\lambda^{n+\frac{1}{2}} \right) \\ q^{n+1,\alpha} &= q^{n+\frac{1}{2}} + \frac{h}{2} p^{n+1,\alpha} \\ G(q^{n+1,0})p^{n+1,0} &= 0. \end{aligned} \tag{88}$$

Remark 2.14. For the implementation we need to solve the weakly nonlinear equations (89) in $\lambda^{n+\frac{1}{2}}$

$$p^{n+1,0} = p^n - h\nabla V(q^{n+\frac{1}{2}}) - hG^{\mathrm{T}}(q^{n+\frac{1}{2}})\lambda^{n+\frac{1}{2}} \tag{89a}$$

$$hG\left(q^{n+\frac{1}{2}} + \frac{h}{2}p^{n+1,0}\right)G^{\mathrm{T}}(q^{n+\frac{1}{2}})\,\lambda^{n+\frac{1}{2}} = G\left(q^{n+\frac{1}{2}} + \frac{h}{2}p^{n+1,0}\right)\left(p^n - h\nabla V(q^{n+\frac{1}{2}})\right) \tag{89b}$$

Even though our method does not preserve the constraint exactly, we can conclude at least the following consistency results.

Lemma 2.15. *Let $\alpha = 0$ and $\kappa \in \mathbb{N}$. Let (q^κ, p^κ) be the numerical solution given by applying method (88) κ – times to initial data $(q^0, p^0) \in \mathbb{R}^{2N}$, which satisfy $G(q^0)p^0 = 0$. Then (q^κ, p^κ) is consistent with the analytical solution (q, p) of (60) at time $T = h\kappa$ for initial condition q^0, p^0. More specifically,*

$$\|p^\kappa - p(T)\| \le ch \tag{90}$$

$$\|q^\kappa - q(T)\| \le \tilde{c}h^2 \tag{91}$$

$$\|G(q^\kappa)p^\kappa\| = 0 \tag{92}$$

$$\|g(q(T)) - g(q^\kappa)\| \le \hat{c}h \tag{93}$$

where the constants $\tilde{c}, \hat{c}, c$ are independent of h and κ.

Proof. The proof is following [29]. While first order consistency is essentially proven by a classical Taylor expansion argument, one still needs to address the algebraic constraint. At the continuous level this is readily achieved by reference to (9) which becomes

$$\lambda = -\left(G(q)G(q)^{\mathrm{T}}\right)^{-1}G(q)\nabla V(q) + \left(G(q)G(q)^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{L}\sum_{j=1}^{L}\left(p^{\mathrm{T}}e_{i,j}p\right)\,\frac{\partial^2 g(q)}{\partial q_i \partial q_j}. \tag{94}$$

The momentum equation in (8) is then equivalent to

$$\dot{p} = -\mathcal{P}_q^{\perp}\nabla V(q) + G(q)^{\mathrm{T}}(G(q)G(q)^{\mathrm{T}})^{-1}\sum_{i=1}^{L}\sum_{j=1}^{L}\left(p^{\mathrm{T}}e_{i,j}p\right)\,\frac{\partial^2 g(q)}{\partial q_i \partial q_j}. \tag{95}$$

For the discrete case we observe

$$G(q^{n+1,0})p^n = G(q^n)p^n + \sum_{i=0}^{N}\sum_{j=0}^{N}(q^{n+1,0}-q^n)e_{i,j}p^n\frac{\partial^2 g(q^n)}{\partial q_i \partial q_j} + \mathcal{O}(h^2) \tag{96}$$

$$= h\sum_{i=0}^{N}\sum_{j=0}^{N}\frac{p^{n+1,0}+p^n}{2}e_{i,j}p^n\frac{\partial^2 g(q^{n+\frac{1}{2}})}{\partial q_i \partial q_j} + \mathcal{O}(h^2), \tag{97}$$

by Taylor expansion, where $G(q^n)p^n = 0$ holds due to the tangential update of the previous time step or if applicable, by the initial condition in (88). Using this identity and the second and fourth update rules in (88) we can express $h\lambda$ explicitly by

$$h\lambda = \Big(G\left(q^{n+1,0}\right)G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big)^{-1}G\left(q^{n+1,0}\right)\Big(p_n - h\nabla V(q^{n+\frac{1}{2}})\Big) \tag{98}$$

$$= -h\Big(G\left(q^{n+1,0}\right)G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big)^{-1}G\left(q^{n+1,0}\right)\nabla V(q^{n+\frac{1}{2}}) \tag{99}$$

$$+\ h\Big(G\left(q^{n+1,0}\right)G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big)^{-1}\left(\sum_{i=0}^{N}\sum_{j=0}^{N}\frac{p^{n+1,0}+p^n}{2}e_{i,j}p^n\frac{\partial^2 g(q^{n+\frac{1}{2}})}{\partial q_i \partial q_j}\right) + \mathcal{O}(h^2).$$

Expanding by Taylor again, this identity now enables us to rewrite the momentum update to

$$\begin{aligned} p^{n+1,0} = p^n &- h\mathcal{P}^{\perp}_{q^{n+\frac{1}{2}}}\nabla V(q^{n+\frac{1}{2}}) \\ &- hG^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big(G(q^{n+\frac{1}{2}})G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big)^{-1}\sum_{i=0}^{N}\sum_{j=0}^{N}\frac{p^{n+1,0}+p^n}{2}e_{i,j}p^n\frac{\partial^2 g(q^{n+\frac{1}{2}})}{\partial q_i \partial q_j} + \mathcal{O}(h^2). \end{aligned} \tag{100}$$

If we compare the momentum update in closed form (95) and the discretization (100) we immediately observe second order local consistency and therefore first order global consistency

$$\|p^\kappa - p(T)\| \leq ch. \tag{101}$$

To bound the consistency error in the coordinates one combines the update rules in (88) to

$$q^{n+1,0} = q^n + h\frac{p^n + p^{n+1,0}}{2}. \tag{102}$$

Since p_{n+1} this a locally second order consistent approximation we obtain

$$q^{n+1,0} = q^n + h\frac{p(t^n) + p(t^{n+1})}{2} + \mathcal{O}(h^3). \tag{103}$$

The implicit midpoint rule is of local third consistency order and we obtain

$$\|q^{n+1,0} - q(t^{n+1})\| = \mathcal{O}(h^3) \tag{104}$$

and therefore

$$\|q^{\kappa} - q(T)\| \leq \tilde{c}h^2. \tag{105}$$

The bound for the constraint (93) follows then directly by expanding $g(q(T))$ around $g(q^n)$. □

Corollary 2.16. *Let additionally $g(q^0) = 0$, then (q^κ, p^κ) is consistent with* (8).

So far we consider only initial data, which is tangential, i.e., satisfies $G(q)p = 0$. This is of course necessary in the context of consistency, since the underlying model is not well posed otherwise. Nevertheless the proposed usage in data assimilation procedures introduces exactly such initial data. The subsequent two statements will clarify what to expect if we apply method (88) to general initial data while $\alpha = 0$.

Lemma 2.17. *Let $\alpha = 0$. The method given in* (88) *approximates the projection of momentum $\mathcal{P}_q^\perp p$ in the following sense.*

$$\|\mathcal{P}_{q^n}^\perp p^n - p^{n+1,0}\| \leq ch \tag{106}$$

Proof. Using the expression for λ as stated in (89b) and subsequently Taylor expansion we conclude

$$\begin{aligned} p^{n+1,0} &= p^n - G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big(G\left(q^{n+1,0}\right) G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big)^{-1} G\left(q^{n+1}\right) p^n + \mathcal{O}(h) && (107)\\ &= p^n - G^{\mathrm{T}}(q^n)\big(G\left(q^n\right) G^{\mathrm{T}}(q^n)\big)^{-1} G\left(q^n\right) p^n + \mathcal{O}(h) && (108)\\ &= \mathcal{P}_{q^n}^\perp p^n + \mathcal{O}(h). && (109) \end{aligned}$$

□

Corollary 2.18. *Let $\alpha = 0$ and $q^0, p^0 \in \mathbb{R}^N$, then method* (88) *is globally first order consistent to the solution given by* (60) *and initial data q^0 and $\mathcal{P}_{q^0}^\perp p^0$.*

Consequently we now establish a consistency result for the blending method, which will provide a connection between the discrete and the analytical evolution.

Lemma 2.19. *Let $\kappa \in \mathbb{N}$ and let $\alpha = \max(0, 1 - dh)$. Furthermore let (q^κ, p^κ) be the numerical solution given by applying (57) κ-times to initial data $(q^0, p^0) \in \mathbb{R}^{2N}$. Then (q^κ, p^κ) is consistent with the solution (q, p) of the dissipative system (59) at time $T = h\kappa$. More specifically,*

$$\|q^\kappa - q(T)\| \leq ch^2 \tag{110}$$

$$\|p^\kappa - p(T)\| \leq \tilde{c}h \tag{111}$$

where $c, \tilde{c}$ is a constant independent of h and κ.

Proof. We start by expressing λ explicitly and rewriting the momentum update as

$$\begin{aligned} p^{n+1,\alpha} &= p_n - h\nabla V(q^{n+\frac{1}{2}}) - \alpha\, h\varepsilon^{-2} G^{\mathrm{T}}(q^{n+\frac{1}{2}}) K g(q_{n+\frac{1}{2}}) - (1-\alpha)\, hG^{\mathrm{T}}(q^{n+\frac{1}{2}})\lambda && (112)\\ &= p^n - h\nabla V(q^{n+\frac{1}{2}}) - \alpha\, h\varepsilon^{-2} G^{\mathrm{T}}(q^{n+\frac{1}{2}}) K g(q^{n+\frac{1}{2}}) && (113)\\ &\quad - (1-\alpha)\, G^{\mathrm{T}}(q^{n+\frac{1}{2}}) \left(G\left(q^{n+1}\right) G^{\mathrm{T}}(q^{n+\frac{1}{2}})\right)^{-1} G\left(q^{n+1,\alpha}\right) \left(p_n - h\nabla V(q^{n+\frac{1}{2}})\right) \\ &= p^n - h(\tilde{\mathcal{P}}^{\perp}_{q^{n+\frac{1}{2}}} + \alpha\tilde{\mathcal{P}}_{q^{n+\frac{1}{2}}})\nabla V(q^{n+\frac{1}{2}}) - h\alpha G^{\mathrm{T}}(q^{n+\frac{1}{2}}) K g(q^{n+\frac{1}{2}}) + (1-\alpha)\tilde{\mathcal{P}}_{q^{n+\frac{1}{2}}} p^n && (114) \end{aligned}$$

where $\tilde{\mathcal{P}}_{q^{n+\frac{1}{2}}} = G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big(G\left(q^{n+1,\alpha}\right) G^{\mathrm{T}}(q^{n+\frac{1}{2}})\Big)^{-1} G\left(q^{n+1,\alpha}\right)$ satisfies by Taylor expansion $\tilde{\mathcal{P}}_{q_{n+\frac{1}{2}}} = \mathcal{P}_{q_{n+\frac{1}{2}}} + \mathcal{O}(h)$. For sufficiently small h we now can substitute $\alpha = \max(0, 1 - hd)$ by $1 - hd$ since d is a fixed number. Expanding by Taylor we conclude first order global consistency by the form of

$$\begin{aligned} p^{n+1,\alpha} &= p^n - h(\tilde{\mathcal{P}}^{\perp}_{q^{n+\frac{1}{2}}} + (1-hd)\tilde{\mathcal{P}}_{q^{n+\frac{1}{2}}})\nabla V(q^{n+\frac{1}{2}}) - h(1-hd)G^{\mathrm{T}}(q^{n+\frac{1}{2}}) K g(q^{n+\frac{1}{2}}) - hd\tilde{\mathcal{P}}_{q^{n+\frac{1}{2}}} p_n && (115)\\ &= p^n - h\nabla V(q^{n+\frac{1}{2}}) - hG^{\mathrm{T}}(q^{n+\frac{1}{2}}) K g(q^{n+\frac{1}{2}}) - hd\mathcal{P}_{q^{n+\frac{1}{2}}} p^n + \mathcal{O}(h^2). && (116) \end{aligned}$$

The update of the coordinates can be rewritten as

$$q^{n+1} = q^n + h\frac{p^n + p^{n+1}}{2} \tag{117}$$

and therefore the same proof as in Lemma 2.15 leads to the statement. □

Remark 2.20. Although the fact that we have to change α depending on h to achieve convergence to a certain model, may seem odd at first thought, this does not contradict the fact that given an a priori choice of α, discrete solutions of the blended method are consistent approximations to solutions of system (59) for a certain d.

To complete and illustrate the overall picture we collect most of the preceding results and references for the

proof of the following commuting diagram.

Proposition 2.21. *Let $\alpha = \max(0, 1 - hd)$. Let ϕ and ψ be the analytical and numerical flows respectively, with regard to the models as mentioned below, then the diagram in Figure 6 commutes.*

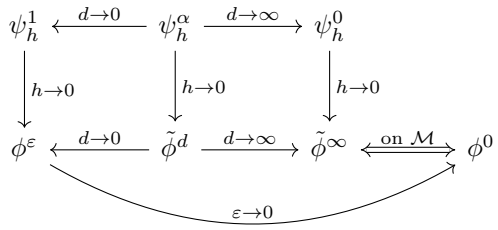

Figure 6: The commuting diagram shows the connections between the analytical model hierarchy given by the flow ϕ^ϵ of the Hamiltonian system (2), the flow $\tilde{\phi}^d$ of the dissipative system (59) and the flow $\tilde{\phi}^\infty$ of the relaxed constrained model (60). Furthermore the diagram depicts the consistency results, of the Störmer-Verlet method (86) and the blended method (57) denoted by ψ_h^α. The flow for the constrained model (8) is mentioned by ψ^0.

Proof. The consistency of the Störmer-Verlet method is stated in e.g. [22] and for an overview of all the other connections in the commuting diagram in Figure 6 we refer to Figure 7. □

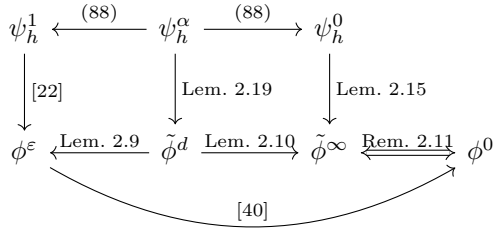

Figure 7: The diagram depicts the same situation as in Figure 6, but refers to the previously established results and relevant literature, instead of the limits.

3 Numerical Results

For experiments in the context of data assimilation one immediate obstacle arises from potential model errors. We avoid this question by considering an initially balanced reference reference solution of (2) which is approximated by the Störmer-Verlet method (86). Henceforth this solution will be denoted by z^{ref}. The observations then are given by $y_{\text{obs}}(t_k) = Hz^{\text{ref}}(t_k) + \zeta_k$. Hereby ζ_k is the realization of the normally distributed measurement error at some time $t_k = k\Delta t_{\text{obs}}$, when the observation becomes available. We assume the measurement error to have zero mean and covariance $R = \rho I$ The resulting evolution of observations is assimilated by the proposed data assimilation

scheme. The advantage of this setup is the straightforward assessment of the quality of the data assimilation method by comparing the reference solution to, e.g., the ensemble members or the point estimate of their mean.

According to (c.f. [39]), due to finite ensemble sizes the true covariances of the posterior distributions are underestimated in ensemble based data assimilation methods. One technique to address this issue is ensemble inflation which amounts to an artificial increase of the spread of the ensemble after each assimilation step by

$$z_i^{\text{new}} := \bar{z} + \sigma_{\text{infl}}(z_i - \bar{z}). \tag{118}$$

In the experiments we apply the ensemble inflation as last step of the assimilation procedure. For the comparison of the presented methods we choose again the stiff elastic double pendulum from Example 1.4 as the dynamical model. The initial ensemble is generated by observations from the initial value of z^{ref}. Subsequently we balance the initial data by solving (25) without the regularization term and for $\gamma = 0$. This is a natural modification of the penalty method for the first step. To compare the methods, we apply this projection to the initial data for every experiment.

For the blended time stepping method we choose a linear ramp for α as depicted in 5 where $\alpha = 0$ initially and $\alpha = 1$ at the end of the blending window. The analysis of (59) is based on linearization and suggests that similarly to the situation of the harmonic oscillator one could find values of α i.e.also the damping d to resolve dynamics close to the aperiodic case. As we do not further investigate the question for optimal α we choose a linear ramp to step through different values of the damping coefficient as brute force approach.

For the numerical values of the parameters of the experiments we refer to Table 1. As with regard to the

B	L	l	ε	K	a_0	Δt	M	Hz	Δt_{obs}	ρ	ρ_0	σ_{infl}	T
$\mathbb{1}$	2	$(1,1)$	0.001	$\text{diag}(1, 0.04)$	9.81	0.001	20	q	0.1	0.05	0.05	1.05	500

Table 1: Parameters for the numerical experiments using the double pendulum model.The model parameters are given by the equilibrium lengths $l \in \mathbb{R}^L$, the scale separation parameter ε, the stiffness matrix K and the gravity a_0. The model is discretized by the Störmer-Verlet method (86) with step width Δt. M denotes the ensemble size, Δt_{obs} the interval between two observations and Hz the observed variable. We choose ρ as covariance of the measurement error, ρ_0 as the initial uncertainty i.e. the covariance of the initial ensemble and ρ_{infl} as the inflation factor. Finally T denotes the duration of the experiment.

implementation details, we solve functional (25) using either the Broyden-Fletcher–Goldfarb-Shanno [10, 17, 19, 41] method as implemented in scipy [43] or the proposed algorithm of (35). For the first we require a tolerance of 10^{-8} and as initial values we choose the results of the plain EnKF. In the second case we choose a fixed step size of $h = 10^{-3}$ and iterate as long as the maximal absolute value of the increment (32) does not exceed 10^{-8}. Additionally we need to solve a nonlinear system for the implicit part of the blended time stepping method (88)

i.e. in the case where $\alpha = 0$. This system is solved using the scipy [43] wrapper for the modified Powell method from the MINPACK [35] subroutine *hybrd*. The initial value is the zero vector of dimension L and the tolerance for the nonlinear problem is set to double precision i.e. 10^{-16}. To quantify the error of the methods we use the time averaged root mean square error as given in [39]

$$\mathrm{TRMSE}(Z) = \sqrt{\frac{1}{N_T}\sum_{i}^{N_T} \|\hat{Z}(t_i) - Z(t_i)\|^2}. \tag{119}$$

Hereby $\hat{Z}$ denotes the estimate for the quantity Z and both are evaluated at N_T time points $t_k \in [0, T]$.

In the following we show this scoring rule in dependence on the tuning parameters of the respective method. For comparison we furthermore show the results for the unmodified ensemble Kalman filter. As one can see in Figures 8 – 10, the forecast quality for the coordinates and the momenta improve drastically when choosing appropriate tuning parameters for the respective methods.

For the penalty method we realize from Figures 8 and 9 that we obtain the best results, when forcing the analysis balance residual of each ensemble member to be close to the respective one inferred from the forecast. We can enforce this by the penalty method when setting $\gamma = 1$. We also find that increased weights do add to the forecast quality only up to certain extent.

The blending method does only allow for one tuning parameter, the windows size. Comparing several choices in Figure 10, we can observe the best results when choosing a window large enough to capture a full period of the less stiff spring in the blending window. This happens approximately around $\frac{2\pi\varepsilon}{\sqrt{k_2}} \approx 0.3\Delta t_{\mathrm{obs}}$.

4 Conclusions

We have proposed two modifications of the standard EnKF when applied to highly oscillatory systems; namely ensemble-based penalization and blended time-stepping. The first one allows flexible use as post processing step for rather generic ensemble based data assimilation algorithms. Both methods perform well in forecast quality and allow accurate state estimation in situations where the standard EnKF fails to do so. The dependency of the forecast skill with regard to the tuning parameters behaves as expected in our prototypical test case of the elastic double pendulum.

We provided a rigorous justification for the blended time-stepping method by the means of asymptotic analysis as well as a numerical method in resemblance of [3]. The optimal choice of the blending window is topic of further investigation.

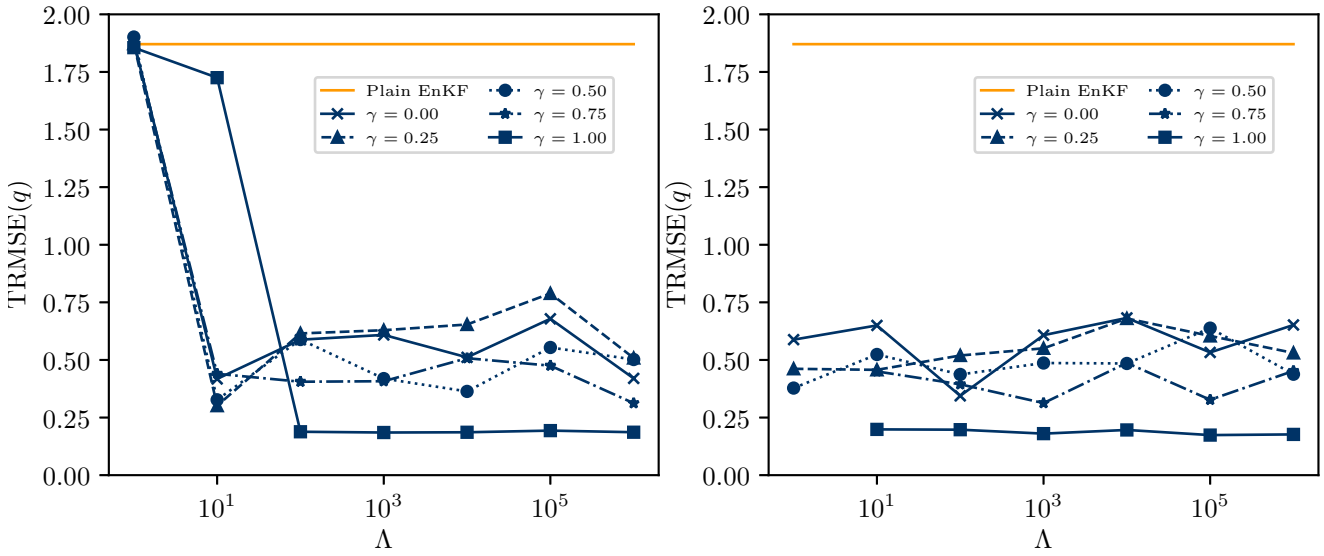

Figure 8: The left figure depicts the time averaged root mean square error (TMRSE) of the coordinates obtained by the penalty method obtained by solving the functional (25) using the previously mentioned BFGS solver. In the right figure we show the results for the same experiment, but using the penalty method solved by (35). In orange we depict the results obtained by the unmodified ensemble Kalman filter.

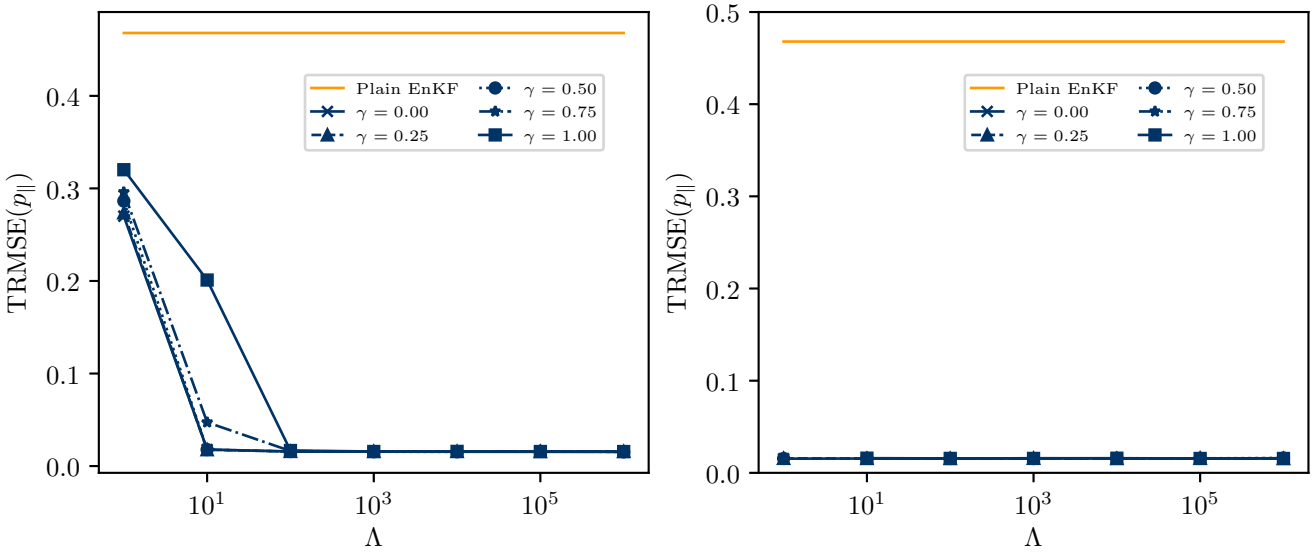

Figure 9: The left and right figures show the time averaged root mean square error in the tangential component of the unobserved momenta, for the penalty method solved by the BFGS and (35) respectively. In orange we depict the results obtained by the unmodified ensemble Kalman filter.

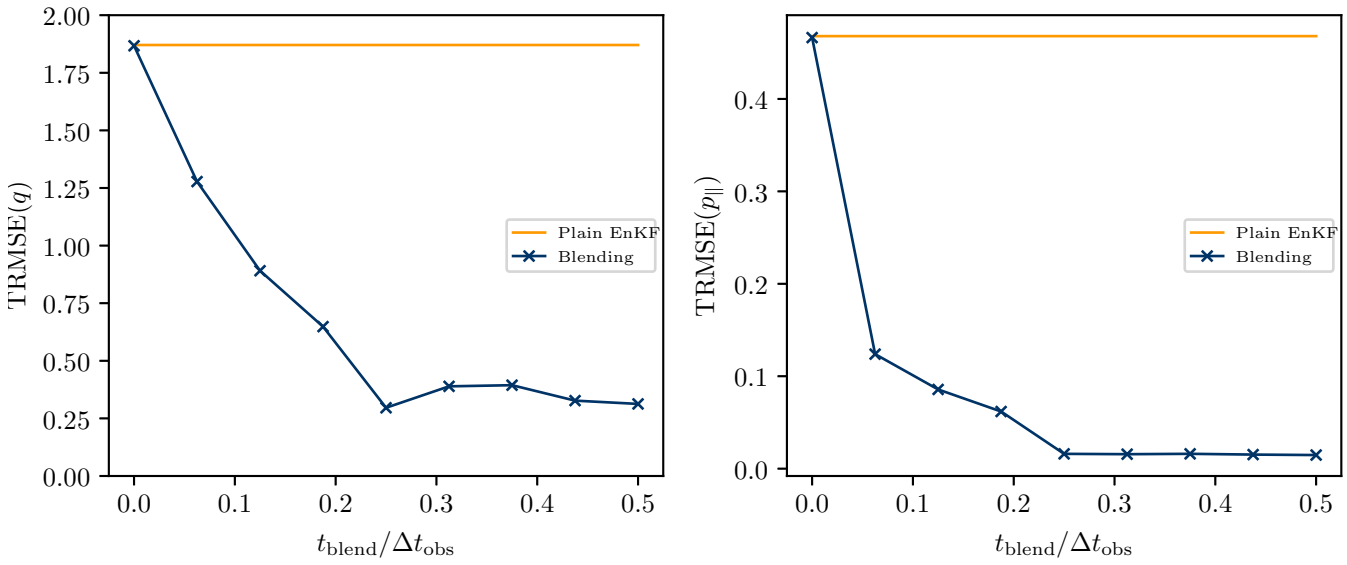

Figure 10: On the left we can see the time averaged root mean square error for the coordinates obtained by the blending method. The right figure shows the same for the tangential component of the unobserved momenta. For comparison the results obtained by the unmodified ensemble Kalman filter are shown in orange. The abscissa describes the ratio of the length of the blending window and the observation interval.

The broader application areas of the two proposed stabilization techniques is ensemble based data assimilation for geophysical processes. This application area shares the situation of small oscillatory energy and conservative motion along a slow manifold. A natural extension of this work will be the investigation of the proposed methods in more realistic geophysical models governed by partial differential equations as e.g. the rotational shallow water equations or the Euler equations. Depending on the scale we can observe several balances in those models. One specific example would be the geostrophic balance. In contrast to the present work, these balances often are linear and therefore Lemma 1.7 actually applies which then implies the weak generation of imbalances. This effect is amplified by the use of localization in the assimilation algorithm which breaks the assumption of a linear filter transformation and therefore gives raise to imbalances again. We emphasize that our methodologies both directly translate to this context, since neither of the algorithms leverages the linearity of the filter. In further publications we will present detailed investigation of the proposed methods applied to a rotational shallow water model as well as a vertical slice model.

Acknowledgments

This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 “Scaling Cascades in Complex Systems”,Project Number 235221301, Project A02 “Multiscale data and asymptotic

model assimilation for atmospheric flows".

References

[1] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, Philadelphia, 1998.

[2] U. M. Ascher, H. Chin, and S. Reich. Stabilization of DAEs and invariant manifolds. *Numerische Mathematik*, 67(2):131–149, 1994.

[3] T. Benacchio, W. P. O'Neill, and R. Klein. A blended soundproof-to-compressible numerical model for small- to mesoscale atmospheric dynamics. *Monthly Weather Review*, 142(12):4416–4438, 2014.

[4] G. Benettin, L. Galgani, and A. Giorgilli. Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory .1. *Communications in Mathematical Physics*, 113(1):87–103, 1987.

[5] G. Benettin, L. Galgani, and A. Giorgilli. Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory .2. *Communications in Mathematical Physics*, 121(4):557–601, 1989.

[6] K. Bergemann and S. Reich. A mollified ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 136:1636–1643, 2010.

[7] S. Bloom, L. L. Takacs, A. Da Silva, and D. Ledvina. Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124:1256–1271, 1996.

[8] O. Bokhove and T. G. Shepherd. On hamiltonian balanced dynamics and the slowest invariant manifold. *Journal of the Atmospheric Sciences*, 53(2):276–297, 1996.

[9] F. A. Bornemann and C. Schütte. Homogenization of Hamiltonian systems with a strong constraining potential. *Physica D*, 102(1-2):57–77, 1997.

[10] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[11] R. Camassa. On the geometry of an atmospheric slow manifold. *Physica D: Nonlinear Phenomena*, 84(3-4): 357–397, 1995.

[12] A. J. Chorin. The numerical solution of Navier-Stokes equations for an imcompressible fluid. *Bulletin of the American Mathematical Society*, 73(6):928–931, 1967.

[13] C. Cotter. Data assimilation on the exponentially accurate slow manifold. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991):20120300, 2013.

[14] J. de Wiljes, S. Reich, and W. Stannat. Long-Time Stability and Accuracy of the Ensemble Kalman–Bucy Filter for Fully Observed Processes and Small Measurement Noise. *SIAM Journal on Applied Dynamical Systems*, 17(2):1152–1181, 2018.

[15] G. Evensen. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343, 2003.

[16] N. Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):53–98, 1979.

[17] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

[18] C. W. Gear. Maintaining solution invariants in the numerical solution of odes. *SIAM J. Sci. Stat. Comput.*, 7 (3):734–743, 1986.

[19] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–23, 1970.

[20] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, 3rd ed edition, 1996.

[21] G. A. Gottwald. Controlling balance in an ensemble Kalman filter. *Nonlinear Processes in Geophysics*, 21(2): 417–426, 2014.

[22] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer series in computational mathematics. Springer, Berlin u.a., 2. edition, 2010.

[23] B. R. Hunt, E. J. Kostelich, and I. Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1-2):112–126, 2007.

[24] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82: 35, 1960.

[25] E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 1 edition, 2002.

[26] J. D. Kepert. Covariance localisation and balance in an Ensemble Kalman Filter. *Quarterly Journal of the Royal Meteorological Society*, 135(642):1157–1176, 2009.

[27] C. Kuehn. *Geometric Singular Perturbation Theory*, pages 53–70. Springer International Publishing, Cham, 2015.

[28] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

[29] B. J. Leimkuhler and R. D. Skeel. Symplectic Numerical Integrators in Constrained Hamiltonian Systems. *Journal of Computational Physics*, 112(1):117–125, 1994.

[30] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.

[31] E. N. Lorenz. Predictability – a problem partly solved. In T. Palmer and R. Hagedorn, editors, *Predictability of Weather and Climate*, page 40–58. Cambridge University Press, 2006.

[32] P. Lynch. The swinging spring: a simple model for atmospheric balance. In J. Norbury and I. Roulstone, editors, *Large-Scale Atmosphere-Ocean Dynamics: Volume II: Geometric Methods and Models*, page 64, 2002.

[33] P. Lynch. *The Emergence of Numerical Weather Prediction: Richardson's Dream*. Cambridge University Press, 2014.

[34] P. Lynch and X.-Y. Huang. Initialization of the HIRLAM model using a digital filter. *Monthly Weather Review*, 120:1019–1034, 1992.

[35] J. J. Moré, B. S. Garbow, and K. E. Hillstrom. User guide for MINPACK-1. Technical Report ANL-80-74, Argonne Nat. Lab., Argonne, IL, 1980.

[36] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2007.

[37] S. Reich. Smoothed dynamics of highly oscillatory Hamiltonian systems. *Physica D*, 89:28—42, 1995.

[38] S. Reich. Smoothed Langevin dynamics of highly oscillatory systems. *Physica D*, 138:210–224, 2000.

[39] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation: a tutorial*. Cambridge University Press, 2015.

[40] H. Rubin and P. Ungar. Motion under a Strong Constraining Force. *Communications on Pure and Applied Mathematics*, 10(1):65–87, 1957.

[41] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–647, 1970.

[42] F. Takens. *Motion under the influence of a strong constraining force*. Springer Berlin Heidelberg, 1980.

[43] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[44] J. S. Whitaker and T. M. Hamill. Ensemble Data Assimilation without Perturbed Observations. *Monthly Weather Review*, 130(7):1913–1924, July 2002. ISSN 0027-0644.

[45] J. Zhou, S. Reich, and B. Brooks. Elastic molecular dynamics with self-consistent flexible constraints. *J. Chem. Phys.*, 112:7919—7929, 2000.