



# Numerical analysis of penalty–based ensemble methods

Rui Fang<sup>1</sup>

Received: 14 August 2024 / Accepted: 14 January 2025

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

## Abstract

The inherent chaos in fluid flow and the uncertainties in initial conditions restrict the ability to make accurate predictions. Small errors that occur in the initial conditions can grow exponentially until they saturate at  $\mathcal{O}(1)$ . Ensemble forecasting averages multiple runs with slightly different initial conditions and other data to produce more accurate results and extend the predictability horizon. However, they can be computationally expensive. We develop a penalty–based ensemble method with a shared coefficient matrix to reduce required memory and computational cost, allowing larger ensemble sizes. Penalty methods relax the incompressibility condition to decouple the pressure and velocity, reducing memory requirements. This report gives stability proof and an error estimate of the penalty–based ensemble method for the Navier–Stokes equations. In addition, we extend the method from deterministic Navier–Stokes equations to Navier–Stokes equations with random variables using Monte Carlo sampling. We validate the method’s accuracy and efficiency with three numerical experiments.

**Keywords** Navier–Stokes equations · Ensemble calculation · Penalty methods · Numerical analysis · Finite element methods

## 1 Introduction

Unstable systems have finite predictability horizons, Lorenz [1, 2]. The chaotic nature of fluid flow and the uncertainties in initial conditions limit predictability. Under different initial conditions, the trajectories of the flow spread. Small errors in the (uncertain) initial conditions can grow exponentially until  $\mathcal{O}(1)$ , resulting in a loss of prediction ability [3].

Ensemble methods address the uncertainty in problem data by conducting numerical simulations with various initial and boundary conditions, external forces, viscosity, and other model parameters, Kalnay [4]. Ensemble mean takes into account the combined information from multiple simulations, effectively smoothing out individual variations and providing a more robust prediction, hence it is considered the best estimate. Leith

---

✉ Rui Fang  
ruf10@pitt.edu

<sup>1</sup> Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260, USA

[5] showed that using a sample size of eight with Monte Carlo sampling can obtain adequate accuracy for the ensemble mean.

Ensemble methods address data uncertainty, thus improving predictability. One remaining issue is that solving  $J$  separate linear systems of equations for an ensemble of size  $J$  can be computationally expensive. Our method resolves this by introducing a shared coefficient matrix, allowing the system to be assembled once and solved efficiently for different right-hand side vectors.

We develop a penalty-based ensemble method for the Navier-Stokes equations (NSE) to reduce computational cost. This enables larger ensemble size, which produces more reliable results. The method uses a shared coefficient matrix with different right-hand side vectors and relaxes the incompressibility condition to reduce the space complexity of the model while maintaining accuracy. Further, eliminating the  $J$  pressure variables saves memory and operations. In this report, we derive a stability proof and an error estimate and conduct three numerical tests to validate the method. In addition, we extend the method from deterministic NSE to NSE with random variables, following an approach similar to that of Luo and Wang [6] for random parabolic partial differential equations.

The incompressible NSE is given by

$$\begin{aligned} \frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u + \nabla p &= f(x, t), \text{ and } \nabla \cdot u = 0, \\ u &= 0 \text{ on the boundary.} \end{aligned} \tag{1.1}$$

where  $u$  denotes the flow velocity and  $p$  denotes the flow pressure. The viscosity is denoted by  $\nu$ , and  $f$  is the body force. In (1.1), the pressure is a Lagrange multiplier to enforce the incompressibility constraint, E and Liu [7]. The penalty method relaxes incompressibility by replacing

$$\nabla \cdot u = 0 \text{ with } \nabla \cdot u^\epsilon + \epsilon p^\epsilon = 0, \text{ for } 0 < \epsilon \ll 1.$$

For the continuous problem, the pressure  $p$  can be replaced by  $p^\epsilon = -\frac{1}{\epsilon} \nabla \cdot u^\epsilon$  to replace  $p$ , which uncouples  $u$  and  $p$  and yields the penalized NSE:

$$\frac{\partial u^\epsilon}{\partial t} + u^\epsilon \cdot \nabla u^\epsilon + \frac{1}{2} (\nabla \cdot u^\epsilon) u^\epsilon - \nu \Delta u^\epsilon - \frac{1}{\epsilon} \nabla \nabla \cdot u^\epsilon = f. \tag{1.2}$$

We adopt the ensemble approach of Nan and Layton [8] to the penalized NSE, using a shared coefficient matrix with different right-hand sides. We suppress the spatial discretization to present the idea. We define the ensemble mean and fluctuation at the timestep  $t_n$ :

$$\langle u^\epsilon \rangle^n := \frac{1}{J} \sum_{j=1}^J u_j^{\epsilon,n}, \text{ and } U_j^{\epsilon,n} := u_j^{\epsilon,n} - \langle u^\epsilon \rangle^n,$$

where  $u_j^{\epsilon,n}$  is the penalized velocity for the  $j^{th}$  ensemble member. We use an implicit-explicit time discretization, which allows the coefficient matrix to be independent of

the ensemble member. The method is to find  $u_j^{\epsilon,n+1}$  in the velocity space and  $p_j^{\epsilon,n+1}$  in the pressure space:

$$\begin{aligned} & \frac{u_j^{\epsilon,n+1} - u_j^{\epsilon,n}}{\Delta t} + \langle u^\epsilon \rangle^n \cdot \nabla u_j^{\epsilon,n+1} + \frac{1}{2}(\nabla \cdot \langle u^\epsilon \rangle^n)u_j^{\epsilon,n+1} \\ & + U_j^{\epsilon,n} \cdot \nabla u_j^{\epsilon,n} + \frac{1}{2}(\nabla \cdot U_j^{\epsilon,n})u_j^{\epsilon,n} - \nu \Delta u_j^{\epsilon,n+1} + \nabla p_j^{\epsilon,n+1} = f_j^{n+1}, \quad (1.3) \\ & \nabla \cdot u_j^{\epsilon,n+1} + \epsilon p_j^{\epsilon,n+1} = 0. \end{aligned}$$

Here,  $\epsilon$  is the same for all ensemble members to ensure a shared coefficient matrix. The ensemble mean drives the flow. We can eliminate the pressure by setting

$$p_j^{\epsilon,n+1} = -\frac{1}{\epsilon} \nabla \cdot u_j^{\epsilon,n+1}$$

to reduce the memory.

### 1.1 Related work

Epstein [9] introduced the first forecasting method that explicitly accounted for the uncertainty in atmospheric model predictions, known as the stochastic–dynamics forecasting method, in 1969. Leith [5] later proposed using ensemble forecasting with multiple members instead of a single realization. He showed that the ensemble mean from Monte Carlo ensembles can achieve accurate results without linear regression. Luo and Wang [6] studied an ensemble algorithm for the deterministic and random parabolic partial differential equations, which led to a single discrete system with multiple right–hand side vectors.

Temam [10] first introduced the penalty method with a modified nonlinear term to ensure energy dissipation. He proved in [10] that  $\lim_{\epsilon \rightarrow 0}(u^\epsilon, p^\epsilon) = (u, p)$ . Penalty methods have been widely studied, including Falk [11], Shen [12] and He [13], He and Li [14]. We can speed up the calculation by eliminating the pressure by  $p^\epsilon = -\frac{1}{\epsilon} \nabla \cdot u^\epsilon$ , Heinrich and Vionnet [15]. The velocity error depends on the penalty parameter  $\epsilon$ , as shown by Bercovier and Engelman [16]. The condition number of the penalized system was studied in Layton and Xu [17], Hughes and Liu and Brooks [18]. Adapting penalty parameters, exploiting  $\epsilon$ –sensitivity, can help with ill–conditioning and provide better accuracy [19–22], and pressure recovery in [23]. Preliminary tests of the penalty–based ensemble method are studied in Fang [24].

## 2 Notations and preliminaries

Let  $D \subset \mathbb{R}^d$  be an open regular domain, where  $d = 2$  or  $3$ . The  $L^2(D)$  norm is denoted as  $\|\cdot\|$ , and the inner product is denoted as  $(\cdot, \cdot)$ . Similarly, we define the  $L^p(D)$  norms  $\|\cdot\|_{L^p}$ , and the Sobolev  $W_p^k(D)$  norms  $\|\cdot\|_{W_p^k}$ . We denote the Sobolev space  $W_2^k(D)$  with norm  $\|\cdot\|_k$  as  $H^k(D)$ . We define the norms for the functions  $v(x, t)$  defined on

$(0, T)$ , for  $1 \leq m < \infty$ ,

$$\|v\|_{\infty,k} := \text{EssSup}_{[0,T]} \|v(t, \cdot)\|_k, \quad \|v\|_{m,k} := \left( \int_0^T \|v(t, \cdot)\|_k^m dt \right)^{1/m}. \quad (2.1)$$

We denote the discrete-time equivalents of the norms as follows:

$$\|v\|_{\infty,k} := \max_{0 \leq n \leq N} \|v^n\|_k, \quad \text{and} \quad \|v\|_{m,k} := \left( \sum_{n=0}^N \|v^n\|_k^m \Delta t \right)^{1/m}. \quad (2.2)$$

Let  $(\Omega, \mathcal{F}, P)$  be a complete probability space, where  $\Omega$  is the set of outcomes,  $\mathcal{F} \subset 2^\Omega$  is the  $\sigma$ -algebra of events, and  $P : \mathcal{F} \rightarrow [0, 1]$  is a probability measure. We denote the set of all integrable functions for the probability measure  $P$  as the  $L^1_P(\Omega)$ . Suppose a random variable  $Y$  such that  $Y \in L^1_P(\Omega)$ , we define the expected value of  $Y$  as follows:

$$E[Y] = \int_{\Omega} Y(\omega) dP(\omega).$$

The stochastic Sobolev spaces are denoted by

$$\tilde{W}_p^k := L^p_P \left( \Omega, W_p^k(D) \right).$$

$\tilde{W}_p^k$  contains stochastic functions  $v : \Omega \times D \rightarrow \mathbb{R}$ , that are measurable with respect to the product  $\sigma$ -algebra  $\mathcal{F} \otimes \mathcal{B}(D)$ , where  $\mathcal{B}$  is a Borel set.  $\tilde{W}_p^k$  is equipped with the averaging norms

$$\|v\|_{\tilde{W}_p^k} = \left( E \left[ \|v\|_{W_p^k(D)}^p \right] \right)^{1/p}.$$

When  $p = 2$ , the above space is a Hilbert space and we have  $\tilde{W}_2^k(D) = \tilde{H}^k(D)$ .

**Lemma 2.1** (See Layton [25], p. 28, p. 29) *Suppose  $\Gamma_0 \subset \partial D$  has a positive measure. Let*

$$H_0^1(D) := \{v \in L^2(D) : \nabla v \in L^2(D) \text{ and } v = 0 \text{ on } \Gamma_0\}. \quad (2.3)$$

*Then, there is a positive constant  $C_{PF}$  such that*

$$\|v\| \leq C_{PF} \|\nabla v\| \text{ for every } v \in H_0^1(D). \quad (2.4)$$

*Thus,  $\|\nabla v\|$  and  $\|v\|$  are equivalent norms on  $H_0^1(D)$ .*

The space  $H^{-k}(D)$  is the dual space of bounded linear functionals on  $H_0^k(D)$ . A norm for  $H^{-1}(D)$  is given by

$$\|f\|_{-1} = \sup_{0 \neq v \in H_0^1(D)} \frac{(f, v)}{\|\nabla v\|}. \quad (2.5)$$

**Lemma 2.2** (See Layton [25], p. 11 ) *Let  $D \subset \mathbb{R}^2$  or  $\mathbb{R}^3$ . If  $f \in L^2(D)$ , then*

$$\|f\|_{-1} \leq C_{PF} \|f\| < \infty.$$

Let  $X$  be the velocity space and  $Q$  be the pressure space:

$$X := (H_0^1(D))^d, \text{ and } Q := L_0^2(D), \tag{2.6}$$

where

$$L_0^2(D) := \{q \in D \rightarrow \mathbb{R} : \int_D q \, dx = 0, q \in L^2(D)\}.$$

We denote the conforming velocity and pressure finite element spaces as follows:

$$X^h \subset X \text{ and } Q^h \subset Q.$$

We assume that  $(X^h, Q^h)$  satisfies the following approximation properties and the Ladyzhenskaya-Babushka-Brezzi condition ( $LBB^h$ ). For  $u \in H^{m+1}(D)^d$  and  $p \in H^m(D)$ ,

$$\begin{aligned} \inf_{v \in X^h} \|\nabla(u - v)\| &\leq Ch^m |u|_{m+1}, \\ \inf_{v \in X^h} \|u - v\| &\leq Ch^{m+1} |u|_{m+1}, \\ \inf_{q \in Q^h} \|p - q\| &\leq Ch^m |p|_m. \end{aligned} \tag{2.7}$$

**Condition 2.3** (See Layton [25], p. 62,  $LBB^h$  condition) *Suppose  $(X^h, Q^h)$  satisfies*

$$\inf_{q^h \in Q^h} \sup_{v_h \in X^h} \frac{(q^h, \nabla \cdot v_h)}{\|v_h\| \|q^h\|} \geq \beta^h > 0, \tag{2.8}$$

where  $\beta^h$  is bounded away from zero uniformly in  $h$ .

Condition 2.3 is equivalent to

$$\beta^h \|q^h\| \leq \sup_{v_h \in X^h} \frac{(q^h, \nabla \cdot v_h)}{\|v_h\|}.$$

We assume the mesh with quasi-uniform triangulation and finite element spaces satisfy the inverse inequality:

$$h \|\nabla v_h\| \leq C \|v_h\| \quad \forall v_h \in X^h. \tag{2.9}$$

**Lemma 2.4** (See Ladyshenskaya [26]) *For any vector function  $u : \mathbb{R}^d \rightarrow \mathbb{R}^d$  with compact support and with finite  $L^p$  norms:*

$$\begin{aligned} \|u\|_{L^4(\mathbb{R}^2)} &\leq 2^{1/4} \|u\|_{L^2(\mathbb{R}^2)}^{1/2} \|\nabla u\|_{L^2(\mathbb{R}^2)}^{1/2}, \quad (d = 2), \\ \|u\|_{L^4(\mathbb{R}^3)} &\leq \frac{4}{3\sqrt{3}} \|u\|^{1/4} \|\nabla u\|^{3/4}, \quad (d = 3), \\ \|u\|_{L^6(\mathbb{R}^3)} &\leq \frac{2}{\sqrt{3}} \|\nabla u\|, \quad (d = 3). \end{aligned} \tag{2.10}$$

**Lemma 2.5** (A discrete Gronwall lemma, see Lemma 5.1, p. 369, [27]) *Let  $\Delta t, B, a_n, b_n, c_n, d_n$  be nonnegative numbers such that for  $l \geq 1$ :*

$$a_l + \Delta t \sum_{n=0}^l b_n \leq \Delta t \sum_{n=0}^{l-1} d_n a_n + \Delta t \sum_{n=0}^l c_n + B, \text{ for } l \geq 0, \tag{2.11}$$

then for all  $\Delta t > 0$ ,

$$a_l + \Delta t \sum_{n=0}^l b_n \leq \exp(\Delta t \sum_{n=0}^{l-1} d_n) (\Delta t \sum_{n=0}^l c_n + B). \tag{2.12}$$

**Lemma 2.6** (See Layton [25], p. 7, Hölder’s and Young’s inequalities) *For any  $\xi > 0, 1 \leq p < \infty$ , and  $\frac{1}{p} + \frac{1}{q} = 1$ , the Hölder and Young’s inequalities:*

$$(u, v) \leq \|u\|_{L^p} \|v\|_{L^q}, \quad (u, v) \leq \frac{\xi}{p} \|u\|_{L^p}^p + \frac{\xi^{-q/p}}{q} \|v\|_{L^q}^q.$$

The generalization for three functions,

$$|fgh| \leq \|f\|_{L^p} \|g\|_{L^q} \|h\|_{L^r}, \text{ where } \frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1. \tag{2.13}$$

The standard skew-symmetric trilinear form is  $\forall u, v, w \in X$ ,

$$b^*(u, v, w) := \frac{1}{2}(u \cdot \nabla v, w) - \frac{1}{2}(u \cdot \nabla w, v).$$

**Lemma 2.7** (See Layton [25], p. 11, Lemma 3) *For any  $u, v, w \in X$ , there is  $C = C(D)$  such that*

$$\begin{aligned} \left| \int_D u \cdot \nabla v \cdot w \, dx \right| &\leq C \|\nabla u\| \|\nabla v\| \|\nabla w\|, \text{ and} \\ \left| \int_D u \cdot \nabla v \cdot w \, dx \right| &\leq C \|u\|^{1/2} \|\nabla u\|^{1/2} \|\nabla v\| \|\nabla w\|. \end{aligned}$$

**Lemma 2.8** (See Layton [25], p. 123, p. 155)  $\forall u, v, w \in X$ ,

$$b^*(u, v, w) = (u \cdot \nabla v, w) + \frac{1}{2} ((\nabla \cdot u)v, w).$$

**Lemma 2.9** (See Layton [25] and Girault and Raviart [28])  $b^*(u, v, w)$  satisfies the following bounds:

$$b^*(u, v, w) \leq \begin{cases} C\sqrt{\|u\|\|\nabla u\|\|\nabla v\|\|\nabla w\|}, \\ C\|\nabla u\|\|\nabla v\|\sqrt{\|w\|\|\nabla w\|}, \\ C\|\nabla u\|\|\nabla v\|\|\nabla w\|. \end{cases} \tag{2.14}$$

for all  $u, v, w \in X$ .

**Definition 2.10**  $P_{Q^h}$  is the  $L^2$  projection of  $Q$  onto  $Q^h$ . That is, for any  $q \in Q$ ,  $P_{Q^h}(q)$  satisfies

$$(P_{Q^h}(q) - q, q^h) = 0, \forall q^h \in Q^h.$$

### 3 Penalty-based ensemble method

We define the final time  $T$  and timestep size at the  $n^{th}$  step  $\Delta t_n$ . The total number of steps  $N$  is given by  $N = T/\Delta t$ . The fully-discrete approximation is then given  $(u_{j,h}^{\epsilon,n}, p_{j,h}^{\epsilon,n}) \in (X^h, Q^h)$ , find  $(u_{j,h}^{\epsilon,n+1}, p_{j,h}^{\epsilon,n+1}) \in (X^h, Q^h)$  satisfying:

$$\begin{aligned} & \frac{1}{\Delta t_n} (u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, v_h) + b^*(\langle u_h^\epsilon \rangle^n, u_{j,h}^{\epsilon,n+1}, v_h) + b^*(u_{j,h}^{\epsilon,n} - \langle u_h^\epsilon \rangle^n, u_{j,h}^{\epsilon,n}, v_h) \\ & + v(\nabla u_{j,h}^{\epsilon,n+1}, \nabla v_h) - (p_{j,h}^{\epsilon,n+1}, \nabla \cdot v_h) + (q^h, \nabla \cdot u_{j,h}^{\epsilon,n+1}) + \epsilon(p_{j,h}^{\epsilon,n+1}, q^h) = (f_j^{n+1}, v_h), \end{aligned} \tag{3.1}$$

for all  $(v_h, q^h) \in (X^h, Q^h)$ .

Due to the stretching term  $b^*(u_{j,h}^{\epsilon,n} - \langle u_h^\epsilon \rangle^n, u_{j,h}^{\epsilon,n}, v_h)$ , we need the CFL timestep restriction:

$$C \frac{\Delta t}{vh} \|\nabla(u_{j,h}^{\epsilon,n} - \langle u_h^\epsilon \rangle^n)\|^2 \leq 1. \tag{3.2}$$

If (3.2) is satisfied, we proceed to the next step. Otherwise, we halve the timestep size and repeat the current step.

The penalty-based ensemble method is a one-step method, which allows us to assume a constant  $\Delta t$  in the following analysis for simplicity. However, the time step can be adapted as needed in numerical tests.

We can replace the pressure with  $p_{j,h}^\epsilon = -\frac{1}{\epsilon} P_{Q^h}(\nabla \cdot u_{j,h}^\epsilon) \in Q^h$ . Consequently, the term  $-(p_{j,h}^\epsilon, \nabla \cdot v^h)$  can be rewritten as  $\frac{1}{\epsilon} (P_{Q^h}(\nabla \cdot u_{j,h}^\epsilon), \nabla \cdot v^h)$ . This reformulation eliminates the pressure variable, reducing computational complexity and speeding up calculations in numerical tests.

### 3.1 Stability

Let the difference between the ensemble member  $j$  and the ensemble average be denoted as

$$U_j^{\epsilon,n} := u_{j,h}^{\epsilon,n} - \langle u_h^\epsilon \rangle^n. \tag{3.3}$$

In Theorem 3.1, we prove the method is nonlinearly and long-time energy stable under the CFL condition:

$$C \frac{\Delta t}{vh} \|\nabla U_j^{\epsilon,n}\|^2 \leq 1.$$

**Theorem 3.1** *Suppose the following timestep condition holds:*

$$C \frac{\Delta t}{vh} \|\nabla U_j^{\epsilon,n}\|^2 \leq 1, \quad j = 1, \dots, J. \tag{3.4}$$

*It yields that for any  $N \geq 1$ :*

$$\begin{aligned} & \frac{1}{2} \|u_{j,h}^{\epsilon,N}\|^2 + \frac{1}{4} \sum_{n=0}^{N-1} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2 + \frac{v\Delta t}{4} \|\nabla u_{j,h}^{\epsilon,N}\|^2 \\ & + \frac{\Delta t}{\epsilon} \sum_{n=0}^{N-1} \|P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1})\|^2 + \frac{v\Delta t}{4} \sum_{n=0}^{N-1} \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 \\ & \leq \frac{\Delta t}{2v} \sum_{n=0}^{N-1} \|f_{j,h}^{n+1}\|_{-1}^2 + \frac{1}{2} \|u_{j,h}^{\epsilon,0}\|^2 + \frac{v\Delta t}{4} \|\nabla u_{j,h}^{\epsilon,0}\|^2. \end{aligned} \tag{3.5}$$

**Proof** Taking the inner product of the momentum equation with  $v_h \in X^h$  yields

$$\begin{aligned} & \frac{1}{\Delta t_n} (u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, v_h) + v(\nabla u_{j,h}^{\epsilon,n+1}, \nabla v_h) + \frac{1}{\epsilon} (P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1}), \nabla \cdot v_h) \\ & + b^*(\langle u_h^\epsilon \rangle^n, u_{j,h}^{\epsilon,n+1}, v_h) + b^*(u_{j,h}^{\epsilon,n} - \langle u_h^\epsilon \rangle^n, u_{j,h}^{\epsilon,n}, v_h) = (f_j^{n+1}, v_h). \end{aligned} \tag{3.6}$$

Set  $v_h = u_{j,h}^{\epsilon,n+1}$ . Multiply  $\Delta t$  to both sides of (3.6) and apply the polarization identity:

$$\begin{aligned} & \frac{1}{2} \|u_{j,h}^{\epsilon,n+1}\|^2 - \frac{1}{2} \|u_{j,h}^{\epsilon,n}\|^2 + \frac{1}{2} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2 + \Delta t b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n+1}) \\ & + v\Delta t \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 + \frac{\Delta t}{\epsilon} \|P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1})\|^2 = \Delta t (f_j^{n+1}, u_{j,h}^{\epsilon,n+1}). \end{aligned} \tag{3.7}$$

Apply Young's inequality to  $(f_j^{n+1}, u_{j,h}^{\epsilon,n+1})$  gives:

$$\begin{aligned} & \frac{1}{2} \|u_{j,h}^{\epsilon,n+1}\|^2 - \frac{1}{2} \|u_{j,h}^{\epsilon,n}\|^2 + \frac{1}{2} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2 + \Delta t b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n+1}) \\ & + v\Delta t \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 + \frac{\Delta t}{\epsilon} \|P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1})\|^2 \leq \frac{v\Delta t}{2} \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 + \frac{\Delta t}{2v} \|f_j^{n+1}\|_{-1}^2. \end{aligned} \tag{3.8}$$

Next, we treat the trilinear term with the help of inverse inequalities and interpolation,

$$\begin{aligned}
 & -\Delta t b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n+1}) = -\Delta t b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}) \\
 & \leq C \Delta t \|\nabla U_j^{\epsilon,n}\| \|\nabla u_{j,h}^{\epsilon,n}\| \left( \|\nabla(u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n})\| \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\| \right)^{1/2} \\
 & \leq C \Delta t \|\nabla U_j^{\epsilon,n}\| \|\nabla u_{j,h}^{\epsilon,n}\| \frac{1}{\sqrt{h}} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\| \\
 & \leq C \frac{\Delta t^2}{h} \|\nabla U_j^{\epsilon,n}\|^2 \|\nabla u_{j,h}^{\epsilon,n}\|^2 + \frac{1}{4} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2.
 \end{aligned} \tag{3.9}$$

Combine terms, we have

$$\begin{aligned}
 & \frac{1}{2} \|u_{j,h}^{\epsilon,n+1}\|^2 - \frac{1}{2} \|u_{j,h}^{\epsilon,n}\|^2 + \frac{1}{4} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2 + \frac{\nu \Delta t}{2} \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 \\
 & + \frac{\Delta t}{\epsilon} \|P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1})\|^2 \leq \frac{\Delta t}{2\nu} \|f_j\|_{-1}^2 + C \frac{\Delta t^2}{h} \|\nabla U_j^{\epsilon,n}\|^2 \|\nabla u_{j,h}^{\epsilon,n}\|^2.
 \end{aligned} \tag{3.10}$$

Add and subtract  $\frac{\nu \Delta t}{4} \|\nabla u_{j,h}^{\epsilon,n}\|^2$ , we have

$$\begin{aligned}
 & \frac{1}{2} \|u_{j,h}^{\epsilon,n+1}\|^2 - \frac{1}{2} \|u_{j,h}^{\epsilon,n}\|^2 + \frac{1}{4} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2 + \frac{\nu \Delta t}{4} \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 \\
 & + \frac{\nu \Delta t}{4} \left( \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 - \|\nabla u_{j,h}^{\epsilon,n}\|^2 \right) + \frac{\Delta t}{\epsilon} \|P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1})\|^2 \\
 & + \frac{\nu \Delta t}{4} \left( 1 - \frac{C \Delta t}{h} \|\nabla U_j^{\epsilon,n}\|^2 \right) \|\nabla u_{j,h}^{\epsilon,n}\|^2 \leq \frac{\Delta t}{2\nu} \|f_j^{n+1}\|_{-1}^2.
 \end{aligned} \tag{3.11}$$

With the CFL condition in (3.4), (3.11) reduces to:

$$\begin{aligned}
 & \frac{1}{2} \|u_{j,h}^{\epsilon,n+1}\|^2 - \frac{1}{2} \|u_{j,h}^{\epsilon,n}\|^2 + \frac{1}{4} \|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2 + \frac{\nu \Delta t}{4} (\|\nabla u_{j,h}^{\epsilon,n+1}\|^2 - \|\nabla u_{j,h}^{\epsilon,n}\|^2) \\
 & + \frac{\Delta t}{\epsilon} \|P_{Q^h}(\nabla \cdot u_{j,h}^{\epsilon,n+1})\|^2 + \frac{\nu \Delta t}{4} \|\nabla u_{j,h}^{\epsilon,n+1}\|^2 \leq \frac{\Delta t}{2\nu} \|f_j^{n+1}\|_{-1}^2.
 \end{aligned} \tag{3.12}$$

Sum over all  $n$  from 0 to  $N - 1$ , we have the final result. □

**Proposition 3.2** Under the CFL condition given in (3.4), the following holds:

$$\Delta t \sum_{n=0}^N \frac{1}{J} \sum_{j=1}^J \|\nabla U_j^{\epsilon,n}\|^2 < C. \tag{3.13}$$

**Proof** The fluctuation of the flow as defined in Jiang, Kaya, and Layton [29], Definition 2.1, is:

$$\frac{1}{J} \sum_{j=0}^J \|\nabla U_j^{\epsilon,n}\|^2 = \frac{1}{J} \sum_{j=0}^J \|\nabla u_{j,h}^{\epsilon,n}\|^2 - \|\nabla \langle u_h^\epsilon \rangle^n\|^2.$$

Sum from  $n = 0$  to  $n = N$ , and multiply by  $\Delta t$ :

$$\Delta t \sum_{n=0}^N \sum_{j=1}^J \|\nabla U_j^{\epsilon,n}\|^2 = \Delta t \sum_{n=0}^N \frac{1}{J} \sum_{j=1}^J \|\nabla u_{j,h}^{\epsilon,n}\|^2 - \Delta t \sum_{n=0}^N \|\nabla \langle u_h^\epsilon \rangle^n\|^2. \tag{3.14}$$

Since

$$\Delta t \sum_{n=0}^N \|\nabla \langle u_h^\epsilon \rangle^n\|^2 \geq 0, \text{ and } \Delta t \sum_{j=1}^J \sum_{n=0}^N \|\nabla U_j^{\epsilon,n}\|^2 \geq 0,$$

it is sufficient to show that

$$\Delta t \sum_{n=0}^N \frac{1}{J} \sum_{j=1}^J \|\nabla u_{j,h}^{\epsilon,n}\|^2$$

is bounded by a finite number. By Theorem 3.1, for  $j = 1, \dots, J$ , we have

$$\Delta t \sum_{n=0}^N \|\nabla u_{j,h}^{\epsilon,n}\|^2 < C.$$

Thus,

$$\Delta t \sum_{n=0}^N \frac{1}{J} \sum_{j=1}^J \|\nabla u_{j,h}^{\epsilon,n}\|^2 < \infty.$$

□

### 3.2 Error estimates

**Definition 3.3** Define the Stokes projection  $P_s : (X, Q) \rightarrow (X^h, Q^h)$  such that  $P_s(u, p) = (\tilde{u}, \tilde{p})$  satisfies:  $\forall v_h \in X^h$  and  $q^h \in Q^h$ ,

$$\begin{aligned} v(\nabla(u - \tilde{u}), \nabla v_h) - (p - \tilde{p}, \nabla \cdot v_h) &= 0, \\ (\nabla \cdot (u - \tilde{u}), q^h) &= 0. \end{aligned} \tag{3.15}$$

**Proposition 3.4** (See John [30], p. 164, Lemma 4.43) *Let the domain  $D$  be bounded with polyhedral and Lipschitz continuous boundary and  $(u, p) \in (X, Q)$ . Suppose  $LBB^h$  Condition 2.3 holds, then it yields*

$$\begin{aligned} \|\nabla(u - \tilde{u})\| &\leq 2 \left(1 + \frac{1}{\beta^h}\right) \inf_{v_h \in X^h} \|\nabla(u - v_h)\| + \inf_{q^h \in Q^h} \|p - q^h\|, \\ \|p - \tilde{p}\| &\leq \frac{2}{\beta^h} \left\{ \left(1 + \frac{1}{\beta^h}\right) \inf_{v_h \in X^h} \|\nabla(u - v_h)\| + \inf_{q^h \in Q^h} \|p - q^h\| \right\}. \end{aligned} \tag{3.16}$$

Denote the error of the  $j^{th}$  simulation at time  $t_n$ ,  $e_j^{\epsilon,n} := u_j^{\epsilon,n} - u_{j,h}^{\epsilon,n}$ . Here,  $u_j^{\epsilon,n}$  is the solution of the penalized NSE at time  $t_n$  and  $u_{j,h}^{\epsilon,n}$  is the fully discretized solution of penalty-based ensemble method.

**Theorem 3.5** Consider the method in (1.3) and assume the condition in (3.4) holds for all  $n$ :

$$C \frac{\Delta t}{\nu h} \|\nabla U_j^{\epsilon,n}\|^2 \leq 1, \quad j = 1, \dots, J,$$

then there are positive constants  $C$  and  $C_0$  independent of  $h$  and  $\Delta t$  such that:

$$\begin{aligned} & \|e_{j,h}^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|e_{j,h}^{\epsilon,n+1} - e_{j,h}^{\epsilon,n}\|^2 + \Delta t \nu \|\nabla e_{j,h}^{\epsilon,N}\|^2 \\ & + C_0 \Delta t \sum_{n=0}^{N-1} \nu \|\nabla e_{j,h}^{\epsilon,n+1}\|^2 \leq \exp(\alpha) \left\{ \|e_{j,h}^{\epsilon,0}\|^2 + \Delta t \nu \|\nabla e_{j,h}^{\epsilon,0}\|^2 \right. \\ & + h^{2m} C(\nu) T \left( \|u_{j,t}^{\epsilon}\|_{\infty,0}^2 + \frac{1}{\nu^2} \|p_{j,t}^{\epsilon}\|_{\infty,0}^2 \right) + (\Delta t)^3 C(\nu) \|u_{j,t}^{\epsilon}\|_{\infty,0}^2 \\ & \quad + h^{2m} \epsilon \Delta t C(\nu, \beta^h) \left( \|u_{j,t}^{\epsilon}\|_{2,0}^2 + \|p_{j,t}^{\epsilon}\|_{2,0}^2 \right) \\ & \left. + h^{2m} C(\nu) T \left( \|u_j^{\epsilon}\|_{2,0}^2 + \frac{1}{\nu^2} \|p_j^{\epsilon}\|_{2,0}^2 \right) + C(\nu) (\Delta t)^2 \|\nabla u_{j,t}^{\epsilon}\|_{\infty,0}^2 \right\}, \end{aligned} \tag{3.17}$$

where

$$\alpha = C(\nu) \Delta t \sum_{n=0}^{N-1} \|\nabla u_j^{\epsilon,n+1}\|^4.$$

**Proof** We evaluate the continuous penalty-based NSE (1.2) at time  $t = t_{n+1}$ . For any  $v_h \in X^h$ , and  $q^h \in Q^h$ ,

$$\begin{aligned} & \left( \frac{u_j^{\epsilon,n+1} - u_j^{\epsilon,n}}{\Delta t}, v_h \right) + b^*(u_j^{\epsilon,n+1}, u_j^{\epsilon,n+1}, v_h) + \nu (\nabla u_j^{\epsilon,n+1}, \nabla v_h) \\ & - (p_j^{\epsilon,n+1}, \nabla \cdot v_h) + (\nabla \cdot u_j^{\epsilon,n+1}, q^h) + \epsilon (p_j^{\epsilon,n+1}, q^h) = (f_j^{n+1}, v_h) - (r_j^{\epsilon,n+1}, v_h), \end{aligned} \tag{3.18}$$

where

$$r_j^{\epsilon,n+1} = u_{j,t}^{\epsilon,n+1} - \frac{u_j^{\epsilon,n+1} - u_j^{\epsilon,n}}{\Delta t}.$$

Subtract equation (3.1) from (3.18). We have

$$\begin{aligned} & \frac{1}{\Delta t} (e_j^{\epsilon,n+1} - e_j^{\epsilon,n}, v_h) + b^*(u_j^{\epsilon,n+1}, u_j^{\epsilon,n+1}, v_h) - b^*(\langle u_h^{\epsilon} \rangle^n, u_{j,h}^{\epsilon,n+1}, v_h) \\ & - b^*(u_{j,h}^{\epsilon,n} - \langle u_h^{\epsilon} \rangle^n, u_{j,h}^{\epsilon,n}, v_h) + \nu (\nabla e_j^{\epsilon,n+1}, \nabla v_h) - (p_j^{\epsilon,n+1} - p_{j,h}^{\epsilon,n+1}, \nabla \cdot v_h) \\ & \quad + (\nabla \cdot e_j^{\epsilon,n+1}, q^h) + \epsilon (p_j^{\epsilon,n+1} - p_{j,h}^{\epsilon,n+1}, q^h) + (r_j^{\epsilon,n+1}, v_h) = 0. \end{aligned} \tag{3.19}$$

Let  $\tilde{u} \in X^h$  and  $\tilde{q} \in Q^h$ , define  $e_j^{\epsilon,n} = \eta_j^{\epsilon,n} - \phi_{j,h}^{\epsilon,n}$ , where  $\eta_j^{\epsilon,n} := u_j^{\epsilon,n} - \tilde{u}$ ,  $\phi_{j,h}^{\epsilon,n} := u_{j,h}^{\epsilon,n} - \tilde{u}$ .

$$\begin{aligned} & \frac{1}{\Delta t}(\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}, v_h) + \nu(\nabla\phi_{j,h}^{\epsilon,n+1}, \nabla v_h) - (p_{j,h}^{\epsilon,n+1} - \tilde{q}, \nabla \cdot v_h) + (\nabla \cdot \phi_{j,h}^{\epsilon,n+1}, q^h) \\ & + \epsilon(p_{j,h}^{\epsilon,n+1} - \tilde{q}, q^h) = \frac{1}{\Delta t}(\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}, v_h) + \nu(\nabla\eta_j^{\epsilon,n+1}, \nabla v_h) - (p_j^{\epsilon,n+1} - \tilde{q}, \nabla \cdot v_h) \quad (3.20) \\ & \quad + (\nabla \cdot \eta_j^{\epsilon,n+1}, q^h) + \epsilon(p_j^{\epsilon,n+1} - \tilde{q}, q^h) + (r_j^{\epsilon,n+1}, v_h) \\ & \quad + b^*(u_j^{\epsilon,n+1}, u_{j,h}^{\epsilon,n+1}, v_h) - b^*((u_h^\epsilon)^n, u_{j,h}^{\epsilon,n+1}, v_h) - b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, v_h). \end{aligned}$$

Let  $\tilde{u} \in X^h$  and  $\tilde{q} \in Q^h$  satisfy the Stokes projection:

$$\begin{aligned} \nu(\nabla(u_j^{\epsilon,n+1} - \tilde{u}), \nabla v_h) - (p_j^{\epsilon,n+1} - \tilde{q}, \nabla \cdot v_h) &= 0 \text{ for all } v_h \in X^h, \\ (\nabla \cdot (u_j^{\epsilon,n+1} - \tilde{u}), q^h) &= 0 \text{ for all } q^h \in Q^h. \end{aligned}$$

Equation (3.20) is simplified to

$$\begin{aligned} & \frac{1}{\Delta t}(\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}, v_h) + \nu(\nabla\phi_{j,h}^{\epsilon,n+1}, \nabla v_h) - (p_{j,h}^{\epsilon,n+1} - \tilde{q}, \nabla \cdot v_h) + (\nabla \cdot \phi_{j,h}^{\epsilon,n+1}, q^h) \\ & + \epsilon(p_{j,h}^{\epsilon,n+1} - \tilde{q}, q^h) = \frac{1}{\Delta t}(\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}, v_h) + \epsilon(p_j^{\epsilon,n+1} - \tilde{q}, q^h) + (r_j^{\epsilon,n+1}, v_h) \quad (3.21) \\ & \quad + b^*(u_j^{\epsilon,n+1}, u_{j,h}^{\epsilon,n+1}, v_h) - b^*((u_h^\epsilon)^n, u_{j,h}^{\epsilon,n+1}, v_h) - b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, v_h). \end{aligned}$$

Set  $v_h = \phi_{j,h}^{\epsilon,n+1}$  and  $q^h = p_{j,h}^{\epsilon,n+1} - \tilde{q}$ , then apply the polarization identity. We have

$$\begin{aligned} & \frac{1}{2\Delta t}(\|\phi_{j,h}^{\epsilon,n+1}\|^2 - \|\phi_{j,h}^{\epsilon,n}\|^2 + \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2) + \nu\|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + \epsilon\|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 \\ & = \frac{1}{\Delta t}(\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) + \epsilon(p_{j,h}^{\epsilon,n+1} - \tilde{q}, p_{j,h}^{\epsilon,n+1} - \tilde{q}) + (r_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \\ & + b^*(u_j^{\epsilon,n+1}, u_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*((u_h^\epsilon)^n, u_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}). \quad (3.22) \end{aligned}$$

We bound the terms on the right-hand side.

$\frac{1}{\Delta t}(\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1})$  term:

$$\begin{aligned} & \frac{1}{\Delta t}(\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) \leq \left\| \frac{\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}}{\Delta t} \right\|_{-1} \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \\ & \leq C(\nu) \left\| \frac{\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}}{\Delta t} \right\|_{-1}^2 + \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 \\ & \leq C(\nu) \left\| \frac{\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}}{\Delta t} \right\|^2 + \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2. \end{aligned}$$

By the integral form of Taylor’s theorem, we have

$$\eta_j^{\epsilon,n+1} = \eta_j^{\epsilon,n} + \int_{t_n}^{t_{n+1}} \eta_{j,t}^\epsilon ds.$$

Divided by  $\Delta t$  on both sides, and take the  $L^2$  norm on  $D$ ,

$$\begin{aligned} \left\| \frac{\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}}{\Delta t} \right\|^2 &= \int_D \left( \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \eta_{j,t}^\epsilon ds \right)^2 dx \\ &\leq \frac{1}{(\Delta t)^2} \int_D \int_{t_n}^{t_{n+1}} 1 ds \int_{t_n}^{t_{n+1}} |\eta_{j,t}^\epsilon|^2 ds dx \\ &\leq \frac{1}{\Delta t} \int_D \int_{t_n}^{t_{n+1}} |\eta_{j,t}^\epsilon|^2 ds dx. \end{aligned}$$

By Fubini’s theorem, we have

$$\left\| \frac{\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}}{\Delta t} \right\|^2 \leq \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \int_D |\eta_{j,t}^\epsilon|^2 dx ds \leq \max_{t_n \leq t \leq t_{n+1}} \|\eta_{j,t}^\epsilon\|^2.$$

Thus, we have

$$\frac{1}{\Delta t} (\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) \leq C(\nu) \max_{t_n \leq t \leq t_{n+1}} \|\eta_{j,t}^\epsilon\|^2 + \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2.$$

$(r_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1})$  term:

$$\begin{aligned} (r_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) &\leq \|r_j^{\epsilon,n+1}\|_{-1} \|\nabla \phi_{j,h}^{\epsilon,n+1}\| \\ &\leq C(\nu) \|r_j^{\epsilon,n+1}\|_{-1}^2 + \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 \\ &\leq C(\nu) \|r_j^{\epsilon,n+1}\|^2 + \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2. \end{aligned} \tag{3.23}$$

Recall  $r_j^{\epsilon,n+1} = u_{j,t}^{\epsilon,n+1} - \frac{u_j^{\epsilon,n+1} - u_j^{\epsilon,n}}{\Delta t}$ . By the integral form of Taylor’s theorem:

$$\begin{aligned} u_j^{\epsilon,n} &= u_j^{\epsilon,n+1} - \Delta t u_{j,t}^{\epsilon,n+1} - \int_{t_n}^{t_{n+1}} u_{j,tt}^\epsilon (t_n - s) ds, \\ r_j^{\epsilon,n+1} &= \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} u_{j,tt}^\epsilon (s - t_n) ds. \end{aligned} \tag{3.24}$$

$$\begin{aligned}
 \|r_j^{\epsilon,n+1}\|^2 &= \int_D \left( \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} u_{j,tt}^\epsilon(s - t_n) ds \right)^2 dx \\
 &\leq \int_D \left( \int_{t_n}^{t_{n+1}} u_{j,tt}^\epsilon ds \right)^2 dx \leq \int_D \left( \int_{t_n}^{t_{n+1}} |u_{j,tt}^\epsilon| ds \right)^2 dx \quad (3.25) \\
 &\leq \int_D \int_{t_n}^{t_{n+1}} 1 ds \int_{t_n}^{t_{n+1}} |u_{j,tt}^\epsilon|^2 ds dx = \Delta t \int_D \int_{t_n}^{t_{n+1}} |u_{j,tt}^\epsilon|^2 ds dx.
 \end{aligned}$$

By Fubini’s theorem, we have

$$\|r_j^{\epsilon,n+1}\|^2 \leq \Delta t \int_{t_n}^{t_{n+1}} \int_D |u_{j,tt}^\epsilon|^2 dx ds.$$

Hence

$$(r_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \leq C(v)(\Delta t)^2 \int_{t_n}^{t_{n+1}} \int_D |u_{j,tt}^\epsilon|^2 dx ds + \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2. \quad (3.26)$$

$\epsilon(p_{j,h}^{\epsilon,n+1} - \tilde{q}, p_j^{\epsilon,n+1} - \tilde{q})$  term:

$$\epsilon(p_{j,h}^{\epsilon,n+1} - \tilde{q}, p_j^{\epsilon,n+1} - \tilde{q}) \leq \frac{\epsilon}{2} \|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 + \frac{\epsilon}{2} \|p_j^{\epsilon,n+1} - \tilde{q}\|^2.$$

Last we bound the trilinear forms, i.e.  $b^*(\cdot, \cdot, \cdot)$ . Denote

$$A := b^*(u_j^{\epsilon,n+1}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*((u_h^\epsilon)^n, u_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}).$$

First, we add and subtract  $b^*(u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1})$ , and add  $b^*(u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) = 0$ . We have

$$\begin{aligned}
 A &= b^*(u_j^{\epsilon,n+1}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*(u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \\
 &\quad + b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}).
 \end{aligned}$$

Since  $u_j^{\epsilon,n+1} - u_{j,h}^{\epsilon,n+1} = \eta_j^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n+1}$ ,  $u_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n+1} = u_j^{\epsilon,n+1} - \eta_j^{\epsilon,n+1}$ . We have

$$\begin{aligned}
 A &= b^*(u_j^{\epsilon,n+1}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*(u_{j,h}^{\epsilon,n}, u_j^{\epsilon,n+1} - \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \\
 &\quad + b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) \\
 &= b^*(u_j^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) + b^*(u_{j,h}^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \\
 &\quad + b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}).
 \end{aligned}$$

We add and subtract  $b^*(u_j^{\epsilon,n}, u_j^{n+1}, \phi_{j,h}^{\epsilon,n+1})$ ,

$$A = b^*(u_j^{\epsilon,n+1} - u_j^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) + b^*(u_j^{\epsilon,n} - u_{j,h}^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) + b^*(u_{j,h}^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) + b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}). \tag{3.27}$$

Denote

$$A_1 := b^*(u_j^{\epsilon,n+1} - u_j^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}), \quad A_2 := b^*(u_j^{\epsilon,n} - u_{j,h}^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}), \\ A_3 := b^*(u_{j,h}^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}), \quad A_4 := b^*(U_j^{\epsilon,n}, u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}).$$

We estimate  $A_i$ , where  $i = 1, \dots, 4$ , as follows. First, we bound  $A_1$ .

$$A_1 \leq C \|\nabla(u_j^{\epsilon,n+1} - u_j^{\epsilon,n})\| \|\nabla u_j^{\epsilon,n+1}\| \|\nabla \phi_{j,h}^{\epsilon,n+1}\| \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla(u_j^{\epsilon,n+1} - u_j^{\epsilon,n})\|^2 \|\nabla u_j^{\epsilon,n+1}\|^2 \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \Delta t \left( \int_{t_n}^{t_{n+1}} \|\nabla u_{j,t}^{\epsilon,n+1}\|^2 dt \right) \|\nabla u_j^{\epsilon,n+1}\|^2 \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) (\Delta t)^2 \max_{t_n \leq t \leq t_{n+1}} \|\nabla u_{j,t}^{\epsilon,n+1}\|^2 \|\nabla u_j^{\epsilon,n+1}\|^2. \tag{3.28}$$

We bound  $A_2$ .

$$A_2 = b^*(\eta_j^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) - b^*(\phi_{j,h}^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}).$$

$$b^*(\eta_j^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \leq C \|\eta_j^{\epsilon,n}\| \|\nabla^{n+1}\| \|\nabla \phi_{j,h}^{\epsilon,n+1}\| \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla u_j^{\epsilon,n+1}\|^2 \|\nabla \eta_j^{\epsilon,n}\|^2. \tag{3.29}$$

$$-b^*(\phi_{j,h}^{\epsilon,n}, u_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \leq C \sqrt{\|\nabla \phi_{j,h}^{\epsilon,n}\| \|\phi_{j,h}^{\epsilon,n}\|} \|\nabla u_j^{\epsilon,n+1}\| \|\nabla \phi_{j,h}^{\epsilon,n+1}\| \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla \phi_{j,h}^{\epsilon,n}\| \|\phi_{j,h}^{\epsilon,n}\| \|\nabla u_j^{\epsilon,n+1}\|^2 \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + \frac{\nu}{4} \|\nabla \phi_{j,h}^{\epsilon,n}\|^2 + C(\nu) \|\phi_{j,h}^{\epsilon,n}\|^2 \|\nabla u_j^{\epsilon,n+1}\|^4. \tag{3.30}$$

Now we bound  $A_3$ .

$$A_3 = -b^*(\eta_j^{\epsilon,n}, \eta_j^{n+1}, \phi_{j,h}^{\epsilon,n+1}) + b^*(\phi_{j,h}^{\epsilon,n}, \eta_j^{n+1}, \phi_{j,h}^{\epsilon,n+1}) + b^*(u_j^{\epsilon,n}, \eta_j^{n+1}, \phi_{j,h}^{\epsilon,n+1}).$$

$$-b^*(\eta_j^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \leq C \|\nabla \eta_j^{\epsilon,n}\| \|\nabla \eta_j^{\epsilon,n+1}\| \|\nabla \phi_{j,h}^{\epsilon,n+1}\| \\ \leq \frac{\nu}{44} \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla \eta_j^{\epsilon,n}\|^2 \|\nabla \eta_j^{\epsilon,n+1}\|^2. \tag{3.31}$$

$$\begin{aligned}
 b^*(\phi_{j,h}^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) &\leq \sqrt{\|\nabla\phi_{j,h}^{\epsilon,n}\| \|\phi_{j,h}^{\epsilon,n}\|} \|\nabla\eta_j^{\epsilon,n+1}\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla\phi_{j,h}^{\epsilon,n}\| \|\phi_{j,h}^{\epsilon,n}\| \|\nabla\eta_j^{\epsilon,n+1}\|^2 \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + \frac{\nu}{4} \|\nabla\phi_{j,h}^{\epsilon,n}\|^2 + C(\nu) \|\nabla\eta_j^{\epsilon,n+1}\|^4 \|\phi_{j,h}^{\epsilon,n}\|^2.
 \end{aligned}
 \tag{3.32}$$

$$\begin{aligned}
 b^*(u_j^n, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) &\leq C \|\nabla u_j^n\| \|\nabla\eta_j^{\epsilon,n+1}\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla u_j^{\epsilon,n}\|^2 \|\nabla\eta_j^{\epsilon,n+1}\|^2.
 \end{aligned}
 \tag{3.33}$$

Last, we bound  $A_4$ .

$$\begin{aligned}
 A_4 &= b^*(U_j^{\epsilon,n}, u_j^{\epsilon,n+1} - u_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) - b^*(U_j^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) \\
 &\quad + b(U_j^{\epsilon,n}, \eta_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) - b^*(U_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}).
 \end{aligned}$$

$$\begin{aligned}
 b^*(U_j^{\epsilon,n}, u_j^{\epsilon,n+1} - u_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) &\leq C \|\nabla U_j^{\epsilon,n}\| \|\nabla(u_j^{\epsilon,n+1} - u_j^{\epsilon,n})\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla U_j^{\epsilon,n}\|^2 \|\nabla(u_j^{\epsilon,n+1} - u_j^{\epsilon,n})\|^2 \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \Delta t \|\nabla U_j^{\epsilon,n}\|^2 \left(\int_{t_n}^{t_{n+1}} \|\nabla u_{j,t}^{\epsilon,n}\|^2 dt\right) \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) (\Delta t)^2 \|\nabla U_j^{\epsilon,n}\|^2 \max_{t_n \leq t \leq t_{n+1}} \|\nabla u_{j,t}^{\epsilon,n}\|^2.
 \end{aligned}
 \tag{3.34}$$

$$\begin{aligned}
 -b^*(U_j^{\epsilon,n}, \eta_j^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1}) &\leq C \|\nabla U_j^{\epsilon,n}\| \|\nabla\eta_j^{\epsilon,n+1}\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla U_j^{\epsilon,n}\|^2 \|\nabla\eta_j^{\epsilon,n+1}\|^2.
 \end{aligned}
 \tag{3.35}$$

$$\begin{aligned}
 b^*(U_j^{\epsilon,n}, \eta_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) &\leq C \|\nabla U_j^{\epsilon,n}\| \|\nabla\eta_j^{\epsilon,n}\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \\
 &\leq \frac{\nu}{44} \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + C(\nu) \|\nabla U_j^{\epsilon,n}\|^2 \|\nabla\eta_j^{\epsilon,n}\|^2.
 \end{aligned}
 \tag{3.36}$$

$$-b^*(U_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) = b^*(U_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}).
 \tag{3.37}$$

Since  $b^*(u, v, w) + b^*(u, w, v) = 0$ , we have

$$\begin{aligned}
 -b^*(U_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}) &= b^*(U_j^{\epsilon,n}, \phi_{j,h}^{\epsilon,n+1}, \phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}) \\
 &\leq C \|\nabla U_j^{\epsilon,n}\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \sqrt{\|\nabla(\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n})\| \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|} \\
 &\leq C \|\nabla U_j^{\epsilon,n}\| \|\nabla\phi_{j,h}^{\epsilon,n+1}\| \frac{1}{\sqrt{h}} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\| \\
 &\leq \frac{C\Delta t}{h} \|\nabla U_j^{\epsilon,n}\|^2 \|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 + \frac{1}{4\Delta t} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2.
 \end{aligned}
 \tag{3.38}$$

Combine terms,

$$\begin{aligned}
 & \frac{1}{2\Delta t} (\|\phi_{j,h}^{\epsilon,n+1}\|^2 - \|\phi_{j,h}^{\epsilon,n}\|^2 + \frac{1}{2}\|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2) + \left(\frac{\nu}{4} - \frac{C\Delta t}{h}\|\nabla U_j^{\epsilon,n}\|^2\right)\|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 \\
 & \quad + \frac{\nu}{2}(\|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 - \|\nabla\phi_{j,h}^{\epsilon,n}\|^2) + \frac{\epsilon}{2}\|p_j^{\epsilon,n+1} - \tilde{q}\|^2 \\
 & \leq C(\nu)\max_{t_n \leq t \leq t_{n+1}}\|\eta_{j,t}^\epsilon\|^2 + C(\nu)(\Delta t)^2 \int_{t_n}^{t_{n+1}} \int_D |u_{j,tt}^\epsilon|^2 dx ds + \frac{\epsilon}{2}\|p_j^{\epsilon,n+1} - \tilde{q}\|^2 \\
 & \quad + C(\nu)(\Delta t)^2 \max_{t_n \leq t \leq t_{n+1}}\|\nabla u_{j,t}^\epsilon\|^2 \|\nabla u_j^{\epsilon,n+1}\|^2 + C(\nu)\|\nabla u_j^{\epsilon,n+1}\|^2 \|\nabla \eta_j^{\epsilon,n}\|^2 \\
 & + C(\nu)\|\phi_{j,h}^{\epsilon,n}\|^2 \|\nabla u_j^{\epsilon,n+1}\|^4 + C(\nu)\|\nabla \eta_j^{\epsilon,n}\|^2 \|\nabla \eta_j^{\epsilon,n+1}\|^2 + C(\nu)\|\nabla \eta_j^{\epsilon,n+1}\|^4 \|\phi_{j,h}^{\epsilon,n}\|^2 \\
 & \quad + C(\nu)\|\nabla u_j^{\epsilon,n}\|^2 \|\nabla \eta_j^{\epsilon,n+1}\|^2 + C(\nu)(\Delta t)^2 \|\nabla U_j^{\epsilon,n}\|^2 \max_{t_n \leq t \leq t_{n+1}}\|\nabla u_{j,t}^\epsilon\|^2 \\
 & \quad + C(\nu)\|\nabla U_j^{\epsilon,n}\|^2 \|\nabla \eta_j^{\epsilon,n+1}\|^2 + C(\nu)\|\nabla U_j^{\epsilon,n}\|^2 \|\nabla \eta_j^{\epsilon,n}\|^2.
 \end{aligned} \tag{3.39}$$

By the CFL condition, we have

$$\frac{\nu}{4} - \frac{C\Delta t}{h}\|\nabla U_j^{\epsilon,n}\|^2 \geq C_0\nu > 0,$$

for some constant  $C_0 > 0$ .

Recall (3.39), multiply by  $2\Delta t$  and organize terms:

$$\begin{aligned}
 & \|\phi_{j,h}^{\epsilon,n+1}\|^2 - \|\phi_{j,h}^{\epsilon,n}\|^2 + \frac{1}{2}\|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2 + C_0\Delta t\nu\|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 \\
 & \quad + \Delta t \left\{ \nu(\|\nabla\phi_{j,h}^{\epsilon,n+1}\|^2 - \|\nabla\phi_{j,h}^{\epsilon,n}\|^2) + \epsilon\|p_j^{\epsilon,n+1} - \tilde{q}\|^2 \right\} \\
 & \leq \Delta t \left\{ C(\nu) \left( \|\nabla u_j^{\epsilon,n+1}\|^4 + \|\nabla \eta_j^{\epsilon,n+1}\|^4 \right) \|\phi_{j,h}^{\epsilon,n}\|^2 \right. \\
 & + C(\nu)\max_{t_n \leq t \leq t_{n+1}}\|\eta_{j,t}^\epsilon\|^2 + C(\nu)(\Delta t)^2 \int_{t_n}^{t_{n+1}} \int_D |u_{j,tt}^\epsilon|^2 dx ds + \epsilon\|p_j^{\epsilon,n+1} - \tilde{q}\|^2 \\
 & \quad + C(\nu)(\|\nabla \eta_j^{\epsilon,n}\|^2 + \|\nabla u_j^{\epsilon,n}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2)\|\nabla \eta_j^{\epsilon,n+1}\|^2 \\
 & \quad + C(\nu)(\|\nabla u_j^{\epsilon,n+1}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2)\|\nabla \eta_j^{\epsilon,n}\|^2 \\
 & \quad \left. + C(\nu)(\Delta t)^2(\|\nabla u_j^{\epsilon,n+1}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \max_{t_n \leq t \leq t_{n+1}}\|\nabla u_{j,t}^\epsilon\|^2 \right\}.
 \end{aligned} \tag{3.40}$$

Take the sum of (3.40) from  $n = 0$  to  $n = N - 1$ , we have

$$\begin{aligned} & \|\phi_{j,h}^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2 + \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,N}\|^2 + C_0 \sum_{n=0}^{N-1} \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 \\ & + \Delta t \sum_{n=0}^{N-1} \epsilon \|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 \leq \|\phi_{j,h}^{\epsilon,0}\|^2 + \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,0}\|^2 \\ & + \Delta t \left\{ \nu (\|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 - \|\nabla \phi_{j,h}^{\epsilon,n}\|^2) + \epsilon \|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 \right\} \\ & \leq \sum_{n=0}^{N-1} \Delta t \left\{ C(\nu) \left( \|\nabla u_j^{\epsilon,n+1}\|^4 + \|\nabla \eta_j^{\epsilon,n+1}\|^4 \right) \|\phi_{j,h}^{\epsilon,n}\|^2 \right. \\ & + C(\nu) \max_{t_n \leq t \leq t_{n+1}} \|\eta_{j,t}^\epsilon\|^2 + C(\nu) (\Delta t)^2 \int_{t_n}^{t_{n+1}} \int_D |u_{j,t}^\epsilon|^2 dx ds \\ & + \epsilon \|p_j^{\epsilon,n+1} - \tilde{q}\|^2 + C(\nu) (\|\nabla \eta_j^{\epsilon,n}\|^2 + \|\nabla u_j^{\epsilon,n}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \|\nabla \eta_j^{\epsilon,n+1}\|^2 \\ & + C(\nu) (\|\nabla u_j^{\epsilon,n+1}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \|\nabla \eta_j^{\epsilon,n}\|^2 \\ & \left. + C(\nu) (\Delta t)^2 (\|\nabla u_j^{\epsilon,n+1}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \max_{t_n \leq t \leq t_{n+1}} \|\nabla u_{j,t}^\epsilon\|^2 \right\}. \end{aligned}$$

By Lemma 2.5, we have

$$\begin{aligned} & \|\phi_{j,h}^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2 + \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,N}\|^2 + C_0 \sum_{n=0}^{N-1} \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 \\ & + \Delta t \sum_{n=0}^{N-1} \epsilon \|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 \leq \exp \left\{ C(\nu) \Delta t \sum_{n=0}^{N-1} \left( \|\nabla u_j^{\epsilon,n+1}\|^4 + \|\nabla \eta_j^{\epsilon,n+1}\|^4 \right) \right\} \\ & \left\{ \|\phi_{j,h}^{\epsilon,0}\|^2 + \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,0}\|^2 + \Delta t \sum_{n=0}^{N-1} \left( C(\nu) \max_{t_n \leq t \leq t_{n+1}} \|\eta_{j,t}^\epsilon\|^2 \right. \right. \\ & + C(\nu) (\Delta t)^2 \int_{t_n}^{t_{n+1}} \int_D |u_{j,t}^\epsilon|^2 dx ds + \epsilon \|p_j^{\epsilon,n+1} - \tilde{q}\|^2 \\ & + C(\nu) (\|\nabla \eta_j^{\epsilon,n}\|^2 + \|\nabla u_j^{\epsilon,n}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \|\nabla \eta_j^{\epsilon,n+1}\|^2 \\ & + C(\nu) (\|\nabla u_j^{\epsilon,n+1}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \|\nabla \eta_j^{\epsilon,n}\|^2 \\ & \left. \left. + C(\nu) (\Delta t)^2 (\|\nabla u_j^{\epsilon,n+1}\|^2 + \|\nabla U_j^{\epsilon,n}\|^2) \max_{t_n \leq t \leq t_{n+1}} \|\nabla u_{j,t}^\epsilon\|^2 \right) \right\}. \end{aligned}$$

By Proposition 3.2, we conclude that

$$\Delta t \sum_{n=0}^N \|\nabla U_j^{\epsilon,n}\|^2 < C.$$

By Proposition 3.4, we have

$$\begin{aligned} & \|\phi_{j,h}^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2 + \Delta t v \|\nabla \phi_{j,h}^{\epsilon,N}\|^2 + \Delta t \sum_{n=0}^{N-1} \epsilon \|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 \\ & + C_0 \sum_{n=0}^{N-1} \Delta t v \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 \leq \exp(\alpha) \left\{ \|\phi_{j,h}^{\epsilon,0}\|^2 + \Delta t v \|\nabla \phi_{j,h}^{\epsilon,0}\|^2 \right. \\ & + C(v) T \left( \inf_{v_h \in X^h} \|\nabla(u_j^\epsilon - v_h)_t\|_{\infty,0}^2 + \inf_{q^h \in Q^h} \|(p_j^\epsilon - q^h)_t\|_{\infty,0}^2 \right) \\ & \quad + (\Delta t)^3 C(v) \|u_{j,t}^\epsilon\|_{\infty,0}^2 \\ & + \epsilon \Delta t C(v, \beta^h) \left( \inf_{v_h \in X^h} \|\nabla(u_j^\epsilon - v_h)_t\|_{2,0}^2 + \inf_{q^h \in Q^h} \|(p_j^\epsilon - q^h)_t\|_{2,0}^2 \right) \\ & + C(v) \left( \inf_{v_h \in X^h} \|\nabla(u_j^\epsilon - v_h)\|_{\infty,0}^2 + \inf_{q^h \in Q^h} \|p_j^\epsilon - q^h\|_{\infty,0}^2 + \|\nabla u_j^\epsilon\|_{\infty,0}^2 + CT \right) \\ & \quad \left( \inf_{v_h \in X^h} \|\nabla(u_j^\epsilon - v_h)\|_{2,0}^2 + \inf_{q^h \in Q^h} \|p_j^\epsilon - q^h\|_{2,0}^2 \right) \\ & \quad \left. + C(v) (\Delta t)^2 \left( \Delta t \|\nabla u_j^\epsilon\|_{2,0}^2 + C \right) \|\nabla u_{j,t}^\epsilon\|_{\infty,0}^2 \right\}, \end{aligned}$$

where

$$\alpha = C(v) \Delta t \sum_{n=0}^{N-1} \|\nabla u_j^{\epsilon,n+1}\|^4. \tag{3.41}$$

Apply interpolation inequalities in (2.7),

$$\begin{aligned} & \|\phi_{j,h}^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2 + \Delta t v \|\nabla \phi_{j,h}^{\epsilon,N}\|^2 + \Delta t \sum_{n=0}^{N-1} \epsilon \|p_{j,h}^{\epsilon,n+1} - \tilde{q}\|^2 \\ & + C_0 \Delta t \sum_{n=0}^{N-1} v \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 \leq \exp(\alpha) \left\{ \|\phi_{j,h}^{\epsilon,0}\|^2 + \Delta t v \|\nabla \phi_{j,h}^{\epsilon,0}\|^2 \right. \\ & + h^{2m} C(v) T \left( \|u_{j,t}^\epsilon\|_{\infty,0}^2 + \frac{1}{v^2} \|p_{j,t}^\epsilon\|_{\infty,0}^2 \right) + (\Delta t)^3 C(v) \|u_{j,t}^\epsilon\|_{\infty,0}^2 \\ & \quad + h^{2m} \epsilon \Delta t C(v, \beta^h) \left( \|u_{j,t}^\epsilon\|_{2,0}^2 + \|p_{j,t}^\epsilon\|_{2,0}^2 \right) \\ & \quad \left. + h^{2m} C(v) T \left( \|u_j^\epsilon\|_{2,0}^2 + \frac{1}{v^2} \|p_j^\epsilon\|_{2,0}^2 \right) + C(v) (\Delta t)^2 \|\nabla u_{j,t}^\epsilon\|_{\infty,0}^2 \right\} \end{aligned}$$

Recall that  $e_j^{\epsilon,n} = \eta_j^{\epsilon,n} - \phi_{j,h}^{\epsilon,n}$ . Using the triangle inequality, we have

$$\begin{aligned} & \|e_j^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|e_j^{\epsilon,n+1} - e_j^{\epsilon,n}\|^2 + \Delta t \nu \|\nabla e_j^{\epsilon,N}\|^2 + C_0 \Delta t \sum_{n=0}^{N-1} \nu \|\nabla e_j^{\epsilon,n+1}\|^2 \\ \leq & \|\phi_{j,h}^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|\phi_{j,h}^{\epsilon,n+1} - \phi_{j,h}^{\epsilon,n}\|^2 + \Delta t \nu \|\nabla \phi_{j,h}^{\epsilon,N}\|^2 + C_0 \Delta t \sum_{n=0}^{N-1} \nu \|\nabla \phi_{j,h}^{\epsilon,n+1}\|^2 \\ & + \|\eta_j^{\epsilon,N}\|^2 + \frac{1}{2} \sum_{n=0}^{N-1} \|\eta_j^{\epsilon,n+1} - \eta_j^{\epsilon,n}\|^2 + \Delta t \nu \|\nabla \eta_j^{\epsilon,N}\|^2 + C_0 \Delta t \sum_{n=0}^{N-1} \nu \|\nabla \eta_j^{\epsilon,n+1}\|^2. \end{aligned}$$

We complete the proof using the previous bounds for the  $\eta_j^\epsilon$  terms. □

Combining Theorem 3.5 with the result of Shen [12], Theorem 4.1, p. 395, and applying the triangle inequality,

$$\|u_j(t_n) - u_{j,h}^{\epsilon,n}\| \leq \|u_j(t_n) - u_j^\epsilon(t_n)\| + \|u_j^\epsilon(t_n) - u_{j,h}^{\epsilon,n}\|.$$

We have the following corollaries.

**Corollary 3.6** *Assume the regular solutions, under the CFL condition in (3.4), we have the following optimal estimates:*

$$\max_{t_n} \|u_j(t_n) - u_{j,h}^{\epsilon,n}\|^2 + \Delta t \sum_{n=1}^N \|\nabla(u_j(t_n) - u_{j,h}^{\epsilon,n})\|^2 \leq C(u_j, \nu, T)(\epsilon + \Delta t + h^m)^2.$$

**Corollary 3.7** *The error between the average of true solution and the average of penalized finite element approximations is*

$$\|\langle u_{t_n} \rangle - \langle u_h^\epsilon \rangle^n\|^2 \leq C(u_1, \dots, u_J, \nu, T)(\epsilon + \Delta t + h^m)^2.$$

**Proof**

$$\|\langle u(t_n) \rangle - \langle u_h^\epsilon \rangle^n\|^2 = \left\| \frac{1}{J} \sum_{j=1}^J (u_j - u_{j,h}^{\epsilon,n}) \right\|^2 = \left( \frac{1}{J} \right)^2 \left\| \sum_{j=1}^J (u_j - u_{j,h}^{\epsilon,n}) \right\|^2.$$

By the Cauchy Schwarz inequality,

$$\left\| \sum_{j=1}^J (u_j - u_{j,h}^{\epsilon,n}) \right\|^2 \leq J \sum_{j=1}^J \|u_j - u_{j,h}^{\epsilon,n}\|^2.$$

By Corollary 3.6,

$$\sum_{j=1}^J \|u_j - u_{j,h}^{\epsilon,n}\|^2 \leq JC(u_1, \dots, u_J, \nu, T)(\epsilon + \Delta + h^m).$$

Thus,

$$\sum_{j=1}^J \|u_j - u_{j,h}^{\epsilon,n}\|^2 \leq J^2 C(u_1, \dots, u_J, v, T)(\epsilon + \Delta + h^m).$$

Hence, we have

$$\|\langle u(t_n) \rangle - \langle u_h^\epsilon \rangle^n\|^2 \leq C(u_1, \dots, u_J, v, T)(\epsilon + \Delta + h^m).$$

□

### 4 Ensemble-based Monte Carlo forecasting

We consider the NSE with random body forces and initial conditions. We find random functions  $u : \Omega \times \bar{D} \times [0, T] \rightarrow \mathbb{R}^d$ , and  $p : \Omega \times \bar{D} \times [0, T] \rightarrow \mathbb{R}$  satisfy

$$\begin{aligned} \frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u + \nabla p &= f(\omega, x, t), \\ \nabla \cdot u &= 0. \end{aligned} \tag{4.1}$$

We choose a set of random samples for the random body force  $f_j \equiv f(\omega_j, \cdot, \cdot)$ , initial condition  $u_j^0 \equiv u^0(\omega_j, \cdot, \cdot)$  for  $j = 1, \dots, J$ . Note that the corresponding solutions  $u(\omega_j, \cdot, \cdot)$  are independent, identically distributed (i.i.d).

The penalty-based ensemble Monte Carlo is defined as follows. Denote  $u_{j,h}^{\epsilon,n} = u_h^\epsilon(\omega_j, x, t_n)$  and  $p_{j,h}^{\epsilon,n} = p^\epsilon(\omega_j, x, t_n)$ . For the  $j^{th}$  ensemble member and for  $0 \leq n \leq N - 1$ , find  $(u_{j,h}^{\epsilon,n+1}, p_{j,h}^{\epsilon,n+1}) \in (X^h, Q^h)$  satisfying:

$$\begin{aligned} &\frac{1}{\Delta t_n} (u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}, v_h) + b^*((u_h^\epsilon)^n, u_{j,h}^{\epsilon,n+1}, v_h) + b^*(u_{j,h}^{\epsilon,n} - \langle u_h^\epsilon \rangle^n, u_{j,h}^{\epsilon,n}, v_h) \\ &+ \nu(\nabla u_{j,h}^{\epsilon,n+1}, \nabla v_h) - (p_{j,h}^{\epsilon,n+1}, \nabla \cdot v_h) + (q^h, \nabla \cdot u_{j,h}^{\epsilon,n+1}) + \epsilon(p_{j,h}^{\epsilon,n+1}, q^h) = (f_j^{n+1}, v_h), \end{aligned} \tag{4.2}$$

for all  $(v_h, q^h) \in (X^h, Q^h)$ . We approximate  $E[u]$  by the sample average of the penalized NSE

$$\frac{1}{J} \sum_{j=1}^J u_h^\epsilon(\omega_j, \cdot, \cdot).$$

Theorem 3.1 together with the property of expectation leads to the following stability analysis for the finite element solution  $u_{j,h}^{\epsilon,n}$ .

**Theorem 4.1** *Suppose the following timestep condition holds:*

$$C \frac{\Delta t}{\nu h} E[\|\nabla U_j^{\epsilon,n}\|^2] \leq 1, j = 1, \dots, J. \tag{4.3}$$

Then for any  $N \geq 1$ :

$$\begin{aligned} & \frac{1}{2} E[\|u_{j,h}^{\epsilon,N}\|^2] + \frac{1}{4} \sum_{n=0}^{N-1} E[\|u_{j,h}^{\epsilon,n+1} - u_{j,h}^{\epsilon,n}\|^2] + \frac{\nu \Delta t}{4} E[\|\nabla u_{j,h}^{\epsilon,N}\|^2] \\ & + \frac{\Delta t}{\epsilon} \sum_{n=0}^{N-1} E[\|P_{Qh} \nabla \cdot u_{j,h}^{\epsilon,n+1}\|^2] + \frac{\nu \Delta t}{4} \sum_{n=0}^{N-1} E[\|\nabla u_{j,h}^{\epsilon,n+1}\|^2] \tag{4.4} \\ & \leq \frac{\Delta t}{2\nu} \sum_{n=0}^{N-1} E[\|f_{j,h}^{n+1}\|_{-1}^2] + \frac{1}{2} E[\|u_{j,h}^0\|^2] + \frac{\nu \Delta t}{4} E[\|\nabla u_{j,h}^0\|^2] \end{aligned}$$

The fully discrete penalty-based ensemble Monte Carlo approximation is defined to be

$$\Psi_h^n = \frac{1}{J} \sum_{j=1}^J u_{j,h}^{\epsilon,n}.$$

We estimate  $E[u^\epsilon(t_n)] - \Psi_h^n$  in averaged norms. We write

$$E[u^\epsilon(t_n)] - \Psi_h^n = (E[u^\epsilon(t_n)] - E[u_{j,h}^{\epsilon,n}]) + (E[u_{j,h}^{\epsilon,n}] - \Psi_h^n).$$

Since  $u_j^\epsilon$  are i.i.d,  $E[u^\epsilon(t_n)] = E[u_j^\epsilon(t_n)]$ . Thus,

$$E[u^\epsilon(t_n)] - \Psi_h^n = \Gamma_h^n + \Gamma_S^n,$$

where  $\Gamma_h^n = E[u_j(t_n)] - E[u_{j,h}^{\epsilon,n}]$  is the discretization error, and  $\Gamma_S^n = E[u_{j,h}^{\epsilon,n}] - \Psi_h^n$  is the statistical error controls by the ensemble size.

**Theorem 4.2** Assume the condition in (3.4) holds for all  $n$ ,

$$C \frac{\Delta t}{\nu h} E[\|\nabla U_j^{\epsilon,n}\|^2] \leq 1, j = 1, \dots, J, \tag{4.5}$$

then there are positive constant  $C$  and  $C_0$  independent of  $h$  and  $\Delta t$  such that

$$\begin{aligned}
 & E[\|e_{j,h}^{\epsilon,N}\|^2] + \frac{1}{2} \sum_{n=0}^{N-1} E[\|e_{j,h}^{\epsilon,n+1} - e_{j,h}^{\epsilon,n}\|^2] + \Delta t \nu E[\|\nabla e_{j,h}^{\epsilon,N}\|^2] \\
 & + C_0 \Delta t \sum_{n=0}^{N-1} \nu E[\|\nabla e_{j,h}^{\epsilon,n+1}\|^2] \leq \exp(\alpha) \left\{ E[\|e_{j,h}^{\epsilon,0}\|^2] + \Delta t \nu E[\|\nabla e_{j,h}^{\epsilon,0}\|^2] \right. \\
 & + h^{2m} C(\nu) T \left( E[\|u_{j,t}^\epsilon\|_{\infty,0}^2] + \frac{1}{\nu^2} E[\|p_{j,t}^\epsilon\|_{\infty,0}^2] \right) + (\Delta t)^3 C(\nu) E[\|u_{j,t}^\epsilon\|_{\infty,0}^2] \\
 & \quad + h^{2m} \epsilon \Delta t C(\nu, \beta^h) \left( E[\|u_{j,t}^\epsilon\|_{2,0}^2] + E[\|p_{j,t}^\epsilon\|_{2,0}^2] \right) \\
 & \left. + h^{2m} C(\nu) T \left( E[\|u_j^\epsilon\|_{2,0}^2] + \frac{1}{\nu^2} E[\|p_j^\epsilon\|_{2,0}^2] \right) + C(\nu) (\Delta t)^2 E[\|\nabla u_{j,t}^\epsilon\|_{\infty,0}^2] \right\}, \tag{4.6}
 \end{aligned}$$

where

$$\alpha = C(\nu) \Delta t \sum_{n=0}^{N-1} E[\|\nabla u^{\epsilon,n+1}\|^4].$$

**Proof** The conclusion follows Theorem 3.5 after applying the expectation on (3.17).□

**Theorem 4.3** Consider the method in (1.3), assume that  $\forall n$ ,

$$C \frac{\Delta t}{\nu h} E[\|\nabla U_j^{\epsilon,n}\|^2] \leq 1, \quad j = 1, \dots, J. \tag{4.7}$$

Then for any  $N \geq 1$ :

$$\begin{aligned}
 & \frac{1}{2} E[\|\Gamma_S^N\|^2] + \frac{1}{4} \sum_{n=0}^{N-1} E[\|\Gamma_S^{n+1} - \Gamma_S^n\|^2] + \frac{\nu \Delta t}{4} E[\|\nabla \Gamma_S^N\|^2] \\
 & + \frac{\Delta t}{\epsilon} \sum_{n=0}^{N-1} E[\|P_{Q^h} \nabla \cdot \Gamma_S^{n+1}\|^2] + \frac{\nu \Delta t}{4} \sum_{n=0}^{N-1} E[\|\nabla \Gamma_S^{n+1}\|^2] \tag{4.8} \\
 & \leq \frac{1}{J} \left\{ \frac{\Delta t}{2\nu} \sum_{n=0}^{N-1} E[\|f_{j,h}^{n+1}\|_{-1}^2] + \frac{1}{2} E[\|u_{j,h}^0\|^2] + \frac{\nu \Delta t}{4} E[\|\nabla u_{j,h}^0\|^2] \right\}.
 \end{aligned}$$

**Proof** Herein, we present the estimate  $E[\|\nabla \Gamma_S^n\|^2]$ . Define

$$\langle u_{j,h}^{\epsilon,n}, u_{j,h}^{\epsilon,n} \rangle := (\nabla u_{j,h}^{\epsilon,n}, \nabla u_{j,h}^{\epsilon,n}).$$

$$\begin{aligned}
 E[\|\nabla\Gamma_S^n\|^2] &= E\left[\left\langle \frac{1}{J} \sum_{i=1}^J (E[u_{i,h}^{\epsilon,n}] - u_{i,h}^{\epsilon,n}), \frac{1}{J} \sum_{i=1}^J (E[u_{j,h}^{\epsilon,n}] - u_{i,h}^{\epsilon,n}) \right\rangle\right] \\
 &= \frac{1}{J^2} \sum_{i=1}^J \sum_{j=1}^J E[\langle E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n}, E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n} \rangle] \\
 &= \frac{1}{J^2} \sum_{j=1}^J E[\langle E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n}, E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n} \rangle].
 \end{aligned}$$

The last equality is due to the fact  $u_{j,h}^{\epsilon,n}$  for  $j = 1, \dots, J$  are i.i.d.. When  $i \neq j$ , the expectation of  $\langle E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n}, E[u_{i,h}^{\epsilon,n}] - u_{i,h}^{\epsilon,n} \rangle$  is zero. We now expand the quantity  $\langle E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n}, E[u_{j,h}^{\epsilon,n}] - u_{j,h}^{\epsilon,n} \rangle$ . Use the fact that  $E[u_h^{\epsilon,n}] = E[u_{j,h}^{\epsilon,n}]$  and  $E[(u_h^{\epsilon,n})^2] = E[(u_{j,h}^{\epsilon,n})^2]$  to obtain

$$\begin{aligned}
 E[\|\nabla\Gamma_S^n\|^2] &= -\frac{1}{J} \|\nabla E[u_{j,h}^{\epsilon,n}]\|^2 + \frac{1}{J} E[\|\nabla u_{j,h}^{\epsilon,n}\|^2] \\
 &\leq \frac{1}{J} E[\|\nabla u_{j,h}^{\epsilon,n}\|^2].
 \end{aligned}$$

The other terms involving  $E[\|\Gamma_S^N\|^2]$ ,  $E[\|\nabla\Gamma_S^N\|]$  and  $E[\|\Gamma_S^{n+1} - \Gamma_S^n\|]$  can be treated similarly. □

The statistical error from sampling  $\Gamma_S^n$  is  $\mathcal{O}(\frac{1}{\sqrt{J}})$ . Combining Theorem 3.5 with the result of Shen [12], Theorem 4.1, p. 395, and using the triangle inequality, we will have the following corollary.

**Corollary 4.4**

$$\begin{aligned}
 \max_{t_n} E[\|u_j(t_n) - u_{j,h}^{\epsilon,n}\|^2] &+ \Delta t \sum_{n=1}^N E[\|\nabla(u_j(t_n) - u_{j,h}^{\epsilon,n})\|^2] \\
 &\leq C(u_j, v, T)(\epsilon + \Delta t + h^m)^2 \\
 &+ \frac{1}{J} \left\{ \frac{\Delta t}{2\nu} \sum_{n=0}^{N-1} E[\|f_{j,h}^{n+1}\|_{-1}^2] + \frac{1}{2} E[\|u_{j,h}^0\|^2] + \frac{\nu \Delta t}{4} E[\|\nabla u_{j,h}^0\|^2] \right\}.
 \end{aligned}$$

**5 Numerical experiments**

We present the results of three numerical tests to illustrate our theory. In the first test, we calculate the convergence rates using exact solutions with an ensemble size of two. Then, we construct a chaotic Lagrangian flow on a cylinder with perturbed body forces. In the third test, we extend the penalty-based ensemble algorithm to include the Coriolis force for a larger ensemble size for the benchmark test problem of flow

past a cylinder. In these tests, we calculate various flow statistics to evaluate the flow dynamics:

$$\begin{aligned}
 |\text{angular momentum}| &:= \left| \int_D \bar{x} \times \bar{u} \, d\bar{x} \right|, \\
 \text{enstrophy} &:= \frac{1}{2} \nu \|\nabla \times \bar{u}\|^2, \\
 \text{kinetic energy} &:= \frac{1}{2} \|\bar{u}\|^2, \\
 \text{viscous dissipation rate} &:= \nu \|\nabla u\|^2, \\
 \text{numerical dissipation rate from backward Euler (BE)} &:= \frac{1}{\Delta t} (u_n - u_{n-1})^2, \\
 \text{numerical dissipation rate from penalizing incompressibility} &:= \frac{1}{\epsilon} \|\nabla \cdot u\|^2.
 \end{aligned}$$

We use a second–order polynomial to approximate the velocity field in the following tests. The unstructured mesh is generated by GMSH [31].

### 5.1 Test for accuracy

We verify the convergence rates for the method in (3.1) with a test from [32], described as follows. In  $D = (0, 1)^2$ , the exact solution is given by

$$\begin{aligned}
 u(x, y, t) &= (\exp(t) \cos(y), \exp(t) \sin(x))^\top, \\
 p(x, y, t) &= (x - y)(1 + t).
 \end{aligned}$$

We calculate the body force  $f$  by substituting  $u$  and  $p$  in the NSE. We impose the Dirichlet boundary conditions, where  $u_h = u_{true}$  on the boundary. We perturb the initial conditions as follows:

$$u_j(x, y, 0) = (1 + \delta_j)u(x, y, 0), \text{ for } j = 1 \text{ and } 2,$$

**Table 1** The rates of convergence for  $u_1$

g	$\max_{t_n} \ u_1(t_n) - u_{1,h}^{\epsilon,n}\ $	rate	$\sqrt{\Delta t \sum_{n=1}^N \ \nabla(u_1(t_n) - u_{1,h}^{\epsilon,n})\ ^2}$	rate
$(\frac{3}{2})^0 \cdot 27$	0.00358	–	0.01353	–
$(\frac{3}{2})^1 \cdot 27$	0.00169	1.91	0.00639	1.91
$(\frac{3}{2})^2 \cdot 27$	0.00076	1.95	0.0029	1.95
$(\frac{3}{2})^3 \cdot 27$	0.00033	1.98	0.00127	1.98
$(\frac{3}{2})^4 \cdot 27$	0.00015	1.99	0.00057	1.99

where  $\delta_1 = 10^{-3}$  and  $\delta_2 = -10^{-3}$ . We set the kinematic viscosity  $\nu = 1$ , the characteristic velocity of the flow  $U = 1$ , the characteristic length  $L = 1$ , and the Reynolds number  $\mathcal{R}e = \frac{UL}{\nu}$ . To discretize the domain, we choose a sequence of mesh sizes  $h = \frac{1}{g}$ , see Tables 1 and 2. We set  $\Delta t = \frac{h^2}{10}$ ,  $\epsilon = \Delta t$ , and  $T = 1$ . We denote the error as  $e(h) = Ch^\beta$ . We solve the convergence rate  $\beta$  via

$$\beta = \frac{\ln(e(h_1)/e(h_2))}{\ln(h_1/h_2)},$$

at two successive values of  $h$ . Tables 1 and 2 show that the convergence rates of  $u_1$  and  $u_2$  are optimal, consistent with the expected second-order accuracy.

### 5.2 Two rotating small cylinders

We construct a simple 2-dimensional time-periodic flow that exhibits Lagrangian chaos, where the motion of fluid particles becomes chaotic, Aref [33]. Aref’s blinking vortex flow is a model system to study chaotic advection and mixing in fluid flows, introduced by Aref [34, 35], and Aref and Balachandar [36]. The stirring was non-smooth over time, achieved using a point vortex. Herein, we use a cylinder with Dirichlet boundary conditions. The domain is a disk with two smaller obstacles inside (see Fig. 1a). We set the outer circle radius  $r_0 = 1$ , the left inner circle radius  $r_1 = 0.1$ , and the right inner circle radius  $r_2 = 0.1$ ,  $c_1 = \frac{1}{2}$ , and  $c_2 = 0$ . We define the domain:

$$D = \{(x, y) : x^2 + y^2 \leq r_0^2, (x+c_1)^2 + (y-c_2)^2 \geq r_1^2, \text{ and } (x-c_1)^2 + (y-c_2)^2 \geq r_2^2\}.$$

Dirichlet boundary conditions on the left and right circles rotate the flow. Figure 1b shows the amplitude of the left and right circles. The left and right amplitudes are constructed using a periodic modulation of oscillations, which transitions smoothly between active and inactive states. This is achieved through the following steps:

**Table 2** The rates of convergence for  $u_2$

g	$\max_{t_n} \ u_2(t_n) - u_{2,h}^{\epsilon,n}\ $	rate	$\sqrt{\Delta t \sum_{n=1}^N \ \nabla(u_2(t_n) - u_{2,h}^{\epsilon,n})\ ^2}$	rate
$(\frac{3}{2})^0 \cdot 27$	0.00356	–	0.01348	–
$(\frac{3}{2})^1 \cdot 27$	0.00168	1.91	0.00636	1.91
$(\frac{3}{2})^2 \cdot 27$	0.00076	1.95	0.00288	1.95
$(\frac{3}{2})^3 \cdot 27$	0.00033	1.98	0.00126	1.98
$(\frac{3}{2})^4 \cdot 27$	0.00015	1.99	0.00057	1.98

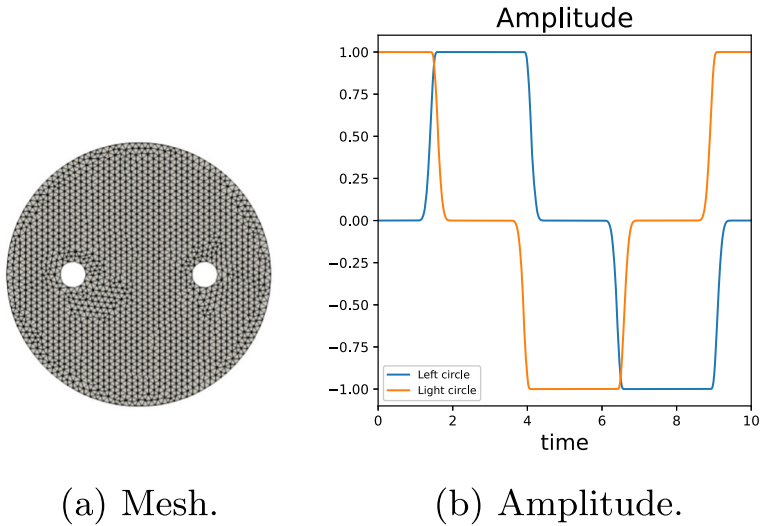


Fig. 1 Mesh configuration and force settings

We define a smooth bridging function,  $\text{smooth\_bridge}(t)$ , which provides a smooth transition near the edges of the active period:

$$\text{smooth\_bridge}(t) = \exp\left(-\frac{\exp\left(-\frac{1}{(1-t)^2}\right)}{t^2}\right), \text{ where } t \in [0, 1].$$

The amplitude function,  $A(t)$ , is periodic with period  $P$  and includes an active phase of length  $L$  within each cycle. It is defined as:

$$A(t) = \begin{cases} 1, & t \bmod P \leq L - 1, \\ \text{smooth\_bridge}(\theta), & L - 1 < t \bmod P \leq L, \text{ with } \theta = L - t \bmod P, \\ 0, & L < t \bmod P \leq P - 1, \\ \text{smooth\_bridge}(1 - \theta), & P - 1 < t \bmod P \leq P. \end{cases}$$

This construction ensures smooth transitions between active and inactive phases. We introduce a larger period,  $P_{\text{large}} = 2P$ , to incorporate alternating positive and negative oscillations. The function is defined as:

$$\text{posAndNegOscillations}(t) = \begin{cases} -A(t) + 1, & t \bmod P_{\text{large}} \leq \frac{P_{\text{large}}}{2}, \\ A(t) - 1, & t \bmod P_{\text{large}} > \frac{P_{\text{large}}}{2}. \end{cases}$$

The left and right amplitudes are derived by phase-shifting the oscillations:

$$A_{\text{left}}(t) = \text{posAndNegOscillations}\left(t + \frac{P}{4} - 1\right),$$

$$A_{\text{right}}(t) = \text{posAndNegOscillations} \left( t + \frac{3P}{4} - 1 \right).$$

We have

$$u(x, y) = 5A_{\text{left/right}}(t)(y, -x)^\top \text{ on } \partial D.$$

We set  $P = 5$  and  $L = 2$ . The outer circle remains stationary. We choose mesh size  $h = 0.05$ , the final time  $T = 10$ , timestep  $\Delta t = 0.001$ ,  $\nu = 1/50$  and  $\mathcal{R}e = 1/\nu$ . The penalty parameter  $\epsilon = \Delta t$ . Flow is at rest at the beginning with exact boundary conditions. We perturbed the Dirichlet boundary conditions:

$$u_{1,2}(x, y) = (1 + \sigma_{1,2})u(x, y) \text{ on } \partial D,$$

where  $\sigma_1 = 0.01$ ,  $\sigma_2 = -0.02$ . For stability, we ensure  $\frac{\Delta t}{h} \|\nabla U_j^{\epsilon, n}\| \leq \frac{10^5}{\mathcal{R}e}$ . This upper bound may not be necessary but sufficient. The solution  $u_0$  is driven by the averaged Dirichlet boundary conditions across the ensemble members:

$$u_0(x, y) = \frac{1}{2}(u_1(x, y) + u_2(x, y)) \text{ on } \partial D.$$

Let  $u_{\text{ave}} = \frac{1}{2}(u_1 + u_2)$ . We define the ensemble spread as:

$$\text{ensemble spread} := \frac{\|u_1 - u_2\|}{\|u_{\text{ave}}\|}.$$

This definition captures the relative difference between the ensemble members normalized by their average. Figure 2a shows that the ensemble spread changes periodically, with the peak of the spread approximately at 0.6. We calculate the standard deviations considering  $u_0$  and  $u_{\text{ave}}$ . Figure 2b shows that the standard deviations for  $u_0$  and  $u_{\text{ave}}$  are similar. It indicates that the velocity is not chaotic. In Fig. 3a, we plot the numerical dissipation rates caused by penalizing the incompressibility condition

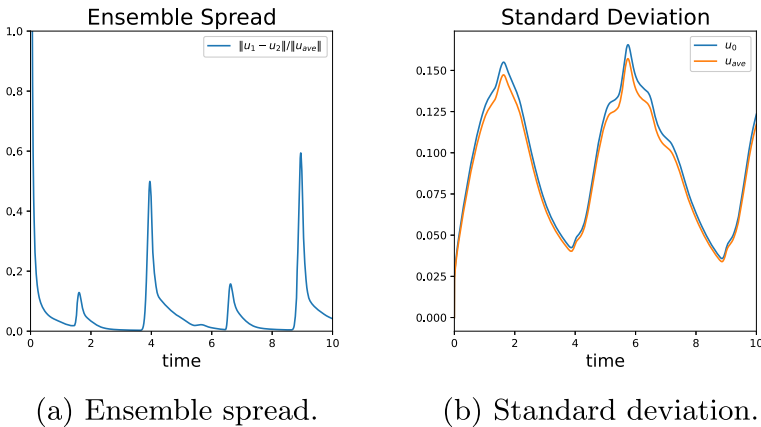


Fig. 2 Ensemble spread and standard deviation

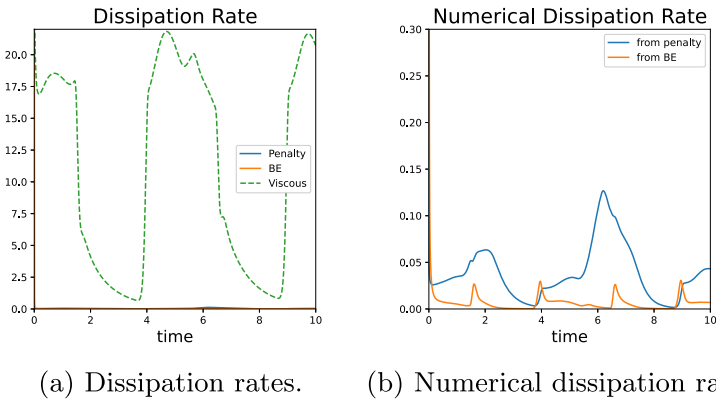


Fig. 3 Dissipation rates

and the BE time discretization. We compare them with the viscous dissipation rate. The numerical dissipation rate is much smaller than the viscous dissipation rate. In Fig. 3b, the numerical dissipation rates have similar magnitudes and vary over time.

We observe changes in kinetic energy, velocity divergence, angular momentum, and enstrophy as we activate and deactivate the spinning of the left and right circles over time. The flow statistics of  $u_0, u_1, u_2,$  and  $u_{ave}$  are closely aligned in Fig. 4 and indicate the velocity field is not chaotic, even though the trajectories of fluid particles exhibit chaotic behavior.

### 5.3 Flow past a cylinder with the Coriolis force for large ensemble sizes

Everything on Earth is rotating even without our noticing. The rotation changes the airflow and affects the climate, as discussed in Lee, Ryi, and Lim [37]. The NSE with the Coriolis force is defined as follows:

$$\frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u + \nabla p + \omega Qu = f.$$

Here  $Q$  is a skew-symmetric matrix with a matrix norm equal to one, and  $\omega$  is the Coriolis coefficient. We extend the penalty-based ensemble method to the NSE with the Coriolis force. We evaluate this method using the benchmark 2D test of flow past a cylinder, as described in [38]. The inlet flow velocity is

$$u(x, y, t) = \left( \frac{6y(0.41 - y)}{0.41^2}, 0 \right)^T.$$

We apply no-slip boundary conditions at the walls and on the obstacle. We generate second-order quadrilateral elements. We choose  $J = 10, T = 10, \Delta t = 0.002, \nu = 0.001,$  and  $\epsilon = \Delta t.$  The flow is at rest at  $t = 0.$  We perturb the inlet flow velocity

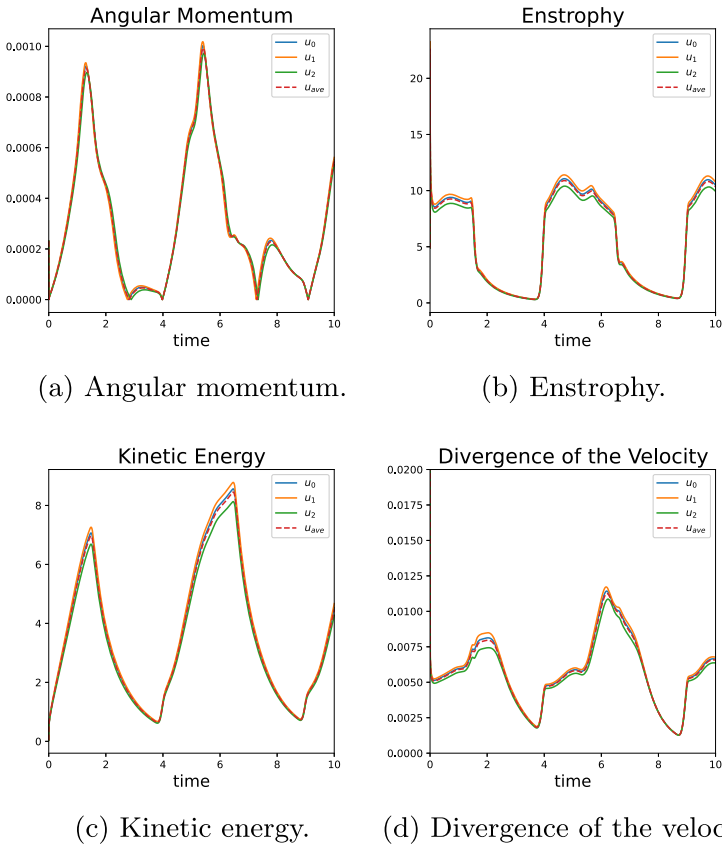


Fig. 4 Flow statistics for  $u_0, u_1, u_2$  and  $u_0$

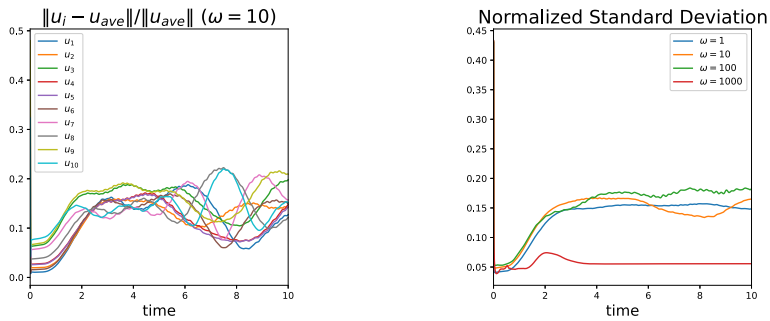


Fig. 5 The normalized standard deviation of the ensembles for different Coriolis coefficients

for ensemble members:

$$u_j(x, y, t) = (1 + \sigma_j \sin(2\pi y))u, \text{ where } j = 1, \dots, 10.$$

Here,  $\sigma_j$  is randomly sampled from  $-0.1$  to  $0.1$ . We first set  $\omega = 10$ . For stability, we ensure  $\frac{\Delta t}{h} \|\nabla U_j^{\epsilon, n}\| \leq \frac{10^5}{Re}$ . Figure 5a shows the spaghetti plot of the relative error of each single ensemble member to the mean flow. The normalized standard deviation for  $\omega = 10$  is around 0.15 after  $t = 2$ , as shown in Fig. 5b. We calculate the angular momentum, enstrophy, kinetic energy, and velocity divergence for all ensemble members and the mean flow, as shown in Fig. 6.

We set the Coriolis coefficient  $\omega = 1, 10, 100,$  and  $1000$  to study the effect of the Coriolis force. We calculate the normalized standard deviation for different values of the Coriolis coefficient, as shown in Fig. 5b. For smaller  $\omega$  values ( $\omega = 1, 10,$  and  $100$ ), the standard deviations are similar, around 0.15. Increasing  $\omega$  to  $1000$  significantly amplifies the rotational force, leading to a smaller standard deviation. Which suggests that the flow exhibits characteristics of rigid body rotation. We also observe much

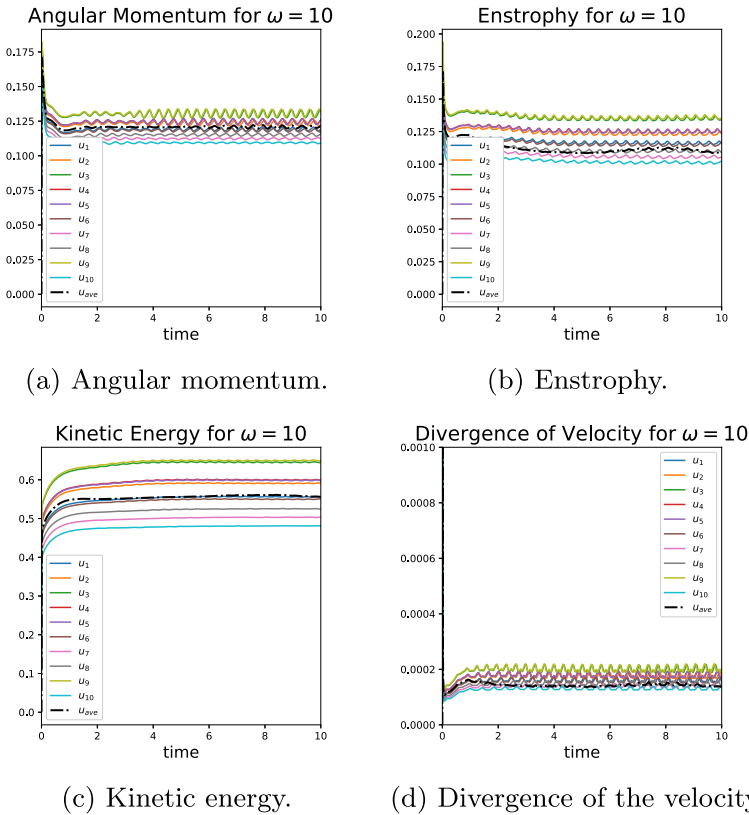


Fig. 6 Flow statistics for all ensemble members and the mean flow at  $\omega = 10$

larger magnitudes of angular momentum, enstrophy, kinetic energy, and divergence of the velocity for the ensemble mean when  $\omega = 1000$ , as shown in Fig. 7.

### 6 Conclusions and prospects

Turbulent flows exhibit chaotic behavior, which inherently limits the predictability of numerical models to a finite horizon. The predictability relies on the accuracy of the initial conditions. Small imperfections in the initial conditions can lead to losing predictive skill. While ensemble methods effectively address this issue, they can be computationally costly. This report introduces a penalty-based ensemble method that reduces the computational cost of ensembles while preserving accuracy. The method uses a shared coefficient matrix for all ensemble members. It relaxes the incompressibility condition, uncoupling the flow velocity and pressure, reducing model complexity, and allowing for a larger ensemble size.

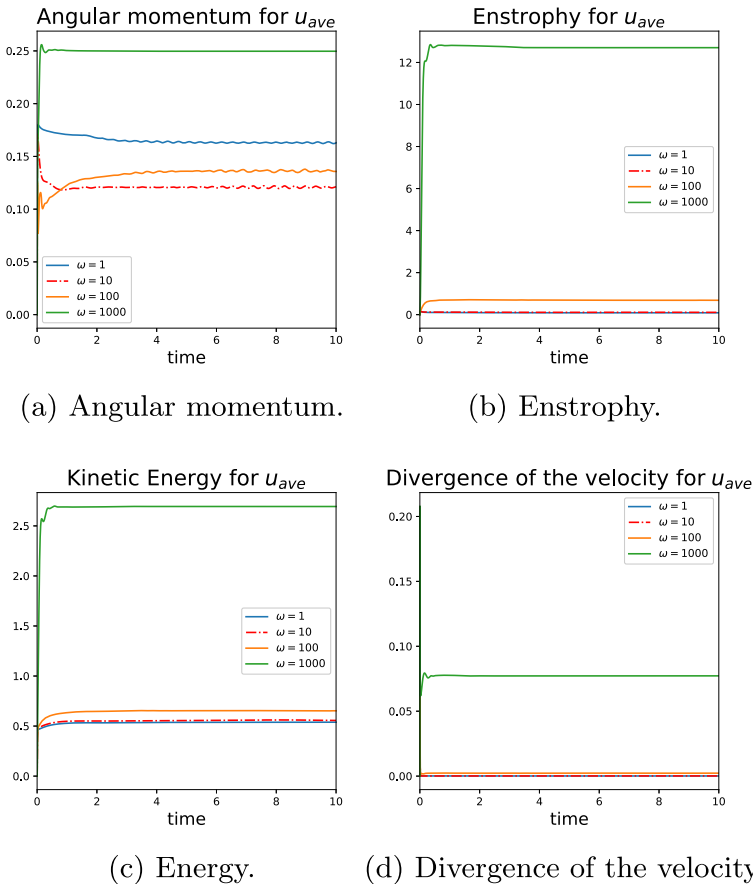


Fig. 7 Flow statistics for the ensemble mean with different Coriolis coefficients

We presented the stability and error estimates of the penalty-based ensemble method in (3.1). We extended the method to the NSE with random body forces and initial conditions with Monte Carlo sampling in Section 4. In Section 5.1, we verified the convergence rates with numerical experiments. In addition, we conducted a numerical experiment on chaotic advection, where the trajectories of the flow particles are chaotic, in Section 5.2. Furthermore, we performed a benchmark test for flow past a cylinder with the Coriolis force using large ensemble sizes in Section 5.3.

Open problems include extending penalty-based ensemble methods to turbulence models [39, 40], adapting penalty parameters for ensemble methods, and exploring alternative techniques for effectively perturbing initial conditions.

**Acknowledgements** I thank my advisor, Professor William Layton, for his guidance and support. We thank Victor DeCaria for a helpful discussion of the test in Section 5.2.

**Author Contributions** R.F. is the sole author and was responsible for writing the manuscript and preparing all figures. R.F. reviewed the manuscript.

**Funding** The research of the author was partially supported by the National Science Foundation under grants DMS-2110379 and DMS-2410893.

**Data Availability** No datasets were generated or analysed during the current study.

## Declarations

**Competing Interests** The authors declare no competing interests.

## References

1. Lorenz, E.N.: Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**(2), 130–141 (1963)
2. Lorenz, E.N.: The predictability of hydrodynamic flow. *Trans. N. Y. Acad. Sci.* **25**(4), 409–432 (1963)
3. Lorenz, E.N.: The growth of errors in prediction. Proceedings of the International School of Physics “Enrico Fermi” Course 88 on Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics. (1985)
4. Kalnay, E.: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, UK (2003)
5. Leith, C.E.: Theoretical skill of monte Carlo forecasts. *Mon. Weather Rev.* **102**(6), 409–418 (1974)
6. Luo, Y., Wang, Z.: An ensemble algorithm for numerical solutions to deterministic and random parabolic PDEs. *SIAM J. Numer. Anal.* **56**(2), 859–876 (2018)
7. E, W., Liu, J.-G.: Projection method I: convergence and numerical boundary layers. *SIAM J. Numer. Anal.* 1017–1057 (1995)
8. Jiang, N., Layton, W.: An algorithm for fast calculation of flow ensembles. *Int. J. Uncertain. Quantif.* **4**(4), (2014)
9. Epstein, E.S.: Stochastic dynamic prediction. *Tellus* **21**(6), 739–759 (1969)
10. Temam, R.: Une méthode d’approximation de la solution des équations de Navier-Stokes. *Bull. Soc. Math. Fr.* **96**, 115–152 (1968)
11. Falk, R.S.: A finite element method for the stationary Stokes equations using trial functions which do not have to satisfy  $\text{div } \mathbf{v} = 0$ . *Math. Comput.* **30**(136), 698–702 (1976)
12. Shen, J.: On error estimates of the penalty method for unsteady Navier-Stokes equations. *SIAM J. Numer. Anal.* **32**(2), 386–403 (1995)
13. He, Y., Li, J.: A penalty finite element method based on the euler implicit/explicit scheme for the time-dependent Navier-Stokes equations. *J. Comput. Appl. Math.* **235**(3), 708–725 (2010)
14. He, Y.: Optimal error estimate of the penalty finite element method for the time-dependent Navier-Stokes equations. *Math. Comput.* **74**(251), 1201–1216 (2005)

15. Heinrich, J., Vionnet, C.A.: The penalty method for the Navier-Stokes equations. *Arch. Comput. Methods Eng.* **2**, 51–65 (1995)
16. Bercovier, M., Engelman, M.: A finite element for the numerical solution of viscous incompressible flows. *J. Comput. Phys.* **30**(2), 181–201 (1979)
17. Layton, W., Xu, S.: Conditioning of linear systems arising from penalty methods. *Electron. Trans. Numer. Anal.* **58**, 394–401 (2023)
18. Hughes, T.J., Liu, W.K., Brooks, A.: Finite element analysis of incompressible viscous flows by the penalty function formulation. *J. Comput. Phys.* **30**(1), 1–60 (1979)
19. Kean, K., Xie, X., Xu, S.: A doubly adaptive penalty method for the Navier Stokes equations. *Int. J. Numer. Anal. Model.* **20**(3), 1 (2023)
20. Xie, X.: On adaptive grad-div parameter selection. *J. Sci. Comput.* **92**(3), 108 (2022)
21. Fang, R.: Numerical analysis of locally adaptive penalty methods for the Navier–Stokes equations. [arXiv:2404.11712](https://arxiv.org/abs/2404.11712) (2024)
22. Layton, W., McLaughlin, M.: Doubly-adaptive artificial compression methods for incompressible flow. *J. Numer. Math.* **28**(3), 175–192 (2020)
23. Kean, K., Schneier, M.: Error analysis of supremizer pressure recovery for pod based reduced-order models of the time-dependent Navier-Stokes equations. *SIAM J. Numer. Anal.* **58**(4), 2235–2264 (2020)
24. Fang, R.: Penalty ensembles for Navier–Stokes with random initial conditions and forcing. [arXiv:2309.12870](https://arxiv.org/abs/2309.12870) (2023)
25. Layton, W.: Introduction to the Numerical Analysis of Incompressible Viscous Flows. SIAM, Philadelphia (2008)
26. Ladyzhenskaya, O.A.: The mathematical theory of viscous incompressible flow. Gordon and Breach (1969)
27. Heywood, J.G., Rannacher, R.: Finite-element approximation of the nonstationary Navier-Stokes problem. Part IV: error analysis for second-order time discretization. *SIAM J. Numer. Anal.* **27**(2), 353–384 (1990)
28. Girault, V., Raviart, P.-A.: Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms, vol. 5. Springer, Berlin, Heidelberg, New York, Tokyo (2012)
29. Jiang, N., Kaya, S., Layton, W.: Analysis of model variance for ensemble based turbulence modeling. *Comput. Methods Appl. Math.* **15**(2), 173–188 (2015)
30. John, V.: Finite Element Methods for Incompressible Flow Problems, vol. 51. Springer, Berlin (2016)
31. Geuzaine, C., Remacle, J.-F.: Gmsh: A 3-D finite element mesh generator with built-in pre-and post-processing facilities. *Int. J. Numer. Methods Eng.* **79**(11), 1309–1331 (2009)
32. Çibik, A., Siddiqua, F., Layton, W.: The ramshaw-mesina hybrid algorithm applied to the Navier Stokes equations. [arXiv:2404.11755](https://arxiv.org/abs/2404.11755) (2024)
33. Aref, H.: Chaotic advection of fluid particles. *Phil. Trans. R. Soc. Lond. Ser. A Phys. Eng. Sci.* **333**(1631), 273–288 (1990)
34. Aref, H.: Stirring by chaotic advection. *J. Fluid Mech.* **143**, 1–21 (1984)
35. Aref, H.: Integrable, chaotic, and turbulent vortex motion in two-dimensional flows. *Annu. Rev. Fluid Mech.* **15**(1), 345–389 (1983)
36. Aref, H., Balachandar, S.: Chaotic advection in a Stokes flow. *Phys. Fluids* **29**(11), 3515–3521 (1986)
37. Lee, S., Ryi, S.-K., Lim, H.: Solutions of Navier–Stokes equation with Coriolis force. *Adv. Math. Phys.* **2017** (2017)
38. Schäfer, M., Turek, S., Durst, F., Krause, E., Rannacher, R.: Benchmark Computations of Laminar Flow Around a Cylinder. Springer, Wiesbaden (1996)
39. Fang, R., Han, W., Layton, W.: On a 1/2–equation model of turbulence. [arXiv:2309.03358](https://arxiv.org/abs/2309.03358) (2023)
40. Han, W.-W., Fang, R., Layton, W.: Numerical analysis of a 1/2-equation model of turbulence. *Phys. D: Nonlinear Phenom.* 134428 (2024)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.