B09209019 葉品辰

1. Introduction

在大氣科學領域中數值模擬和資料同化系統是提升預測能力的關鍵技術，數值模式用於模擬大氣系統的動力演變，而資料同化則結合觀測數據與模式模擬，進一步提升預測的準確性。本研究聚焦於整合 Ensemble Kalman Filter（EnKF）、基於渦度方程式的數值模式，以及機器學習的 U-Net 模型，形成一個創新預測框架。在實驗設計中，為降低運算資源的需求，採用了 U-Net 模型相較於數值模式加速約 1500 倍運算、regrid 技術將高解析度數據降至較小網格，並測試了 Localization、Inflation、系集成員數量及網格大小等參數對同化結果的影響。結果顯示，適當的參數設置可顯著提升系統的準確性與計算效率。本研究未來希望以 U-Net 模型作為高速預測工具驗證其在配合觀測渦度場上資料同化渦度場的潛力，為未來的數值模擬與資料同化應用提供了新的方向。

2. Description of the model and data assimilation systems

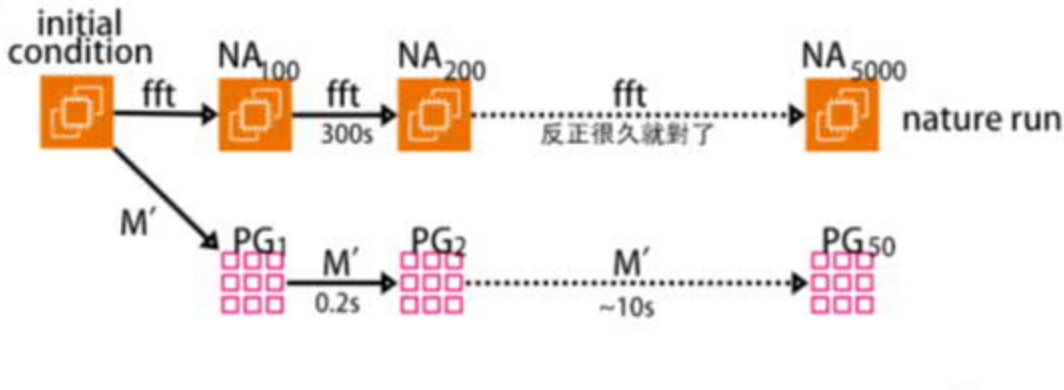
在大氣科學的數值模擬和預測中模型與資料同化系統扮演著關鍵角色。模型用於模擬大氣系統的演變，而資料同化系統則負責結合觀測數據，提升模擬的準確性，並且考慮到真實世界的成本我應該在實驗內限縮算力。以下將介紹本研究中使用的三個主要部分：Ensemble Kalman Filter、數值模式與 U-Net 模型。

A. Ensemble Kalman Filter

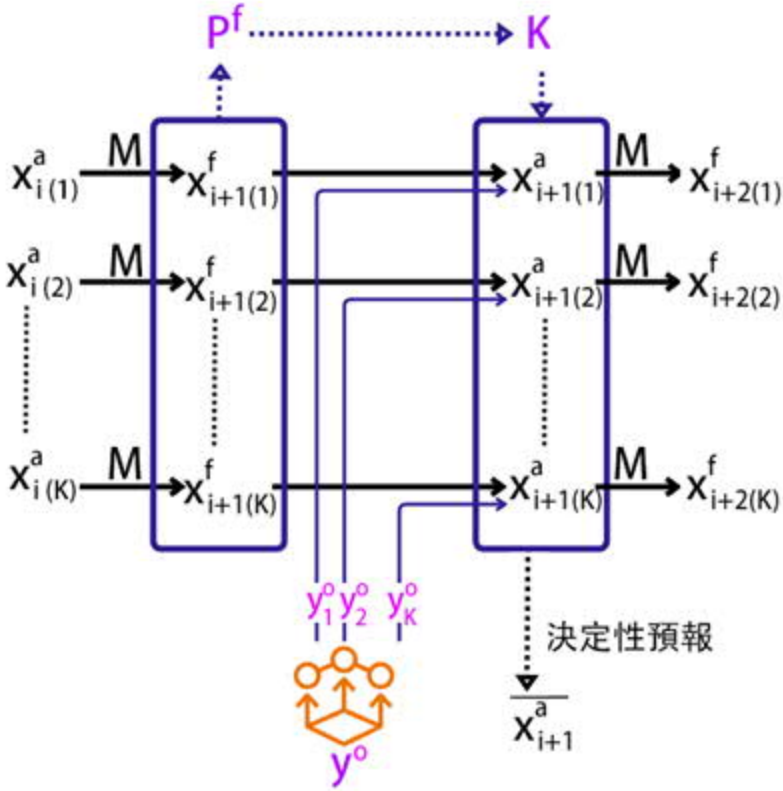
Ensemble Kalman Filter (EnKF) 是一種序列資料同化方法，旨在結合觀測數據與模式模擬來估計動態系統的狀態，它透過一組模擬的集合來表徵系統的不確定性，並利用觀測數據對集合進行更新，以提升預報的準確性。EnKF 可以有效捕捉誤差統計的演變，進而改善系統狀態的估計，預報流程如 fig.1(b)所示、並參考方程組 (1)。

fig.1 model flowchart

(a)



(b)



$$
\begin{cases}
\boldsymbol{P}^f \boldsymbol{H}^T = \frac{1}{K-1}\sum_{k=1}^{K} \boldsymbol{x'}^f_{(k)} \left[H\left(\boldsymbol{x}^f_{(k)} - \overline{H(\boldsymbol{x}^f)}\right)\right]^T & \cdots (1.1) \\
\boldsymbol{H}\boldsymbol{P}^f \boldsymbol{H}^T = \frac{1}{K-1}\sum_{k=1}^{K} \left[H\left(\boldsymbol{x}^f_{(k)} - \overline{H(\boldsymbol{x}^f)}\right)\right]\left[H\left(\boldsymbol{x}^f_{(k)} - \overline{H(\boldsymbol{x}^f)}\right)\right]^T & \cdots (1.2) \\
\boldsymbol{K} = \boldsymbol{P}^f \boldsymbol{H}^T (\boldsymbol{H}\boldsymbol{P}^f \boldsymbol{H}^T + \boldsymbol{R})^{-1} & \cdots (1.3) \\
\boldsymbol{y}^o_{(k)} \sim Normal\ distribution(\boldsymbol{y}^o, \boldsymbol{R}) & \cdots (1.4) \\
\boldsymbol{x}^a_{(k)} = \boldsymbol{x}^f_{(k)} + \boldsymbol{K}\left[\boldsymbol{y}^o_{(k)} - H\left(\boldsymbol{x}^f_{(k)}\right)\right] & \cdots (1.5) \\
\boldsymbol{x}^f_{i+1,(k)} = M_i(\boldsymbol{x}^a_{i,(k)}) & \cdots (1.6)
\end{cases}
$$

本次實驗使用的是 perturbed-observation EnKF (PO-EnKF)，即在進行資料同化時對觀測數據加入小擾動，以生成每一個系集的更新狀態。

B. 數值模式

本研究所使用的數值模式基於渦度方程式，用於模擬初始渦度場的演變，其控制方程式為

$$\frac{\partial \zeta}{\partial t} + \boldsymbol{u} \cdot \nabla\zeta = 0 \quad \cdots (2)$$

其中：ζ 是渦度、 $\boldsymbol{u}$ 是速度場。

該數值模式的空間解析度為 $\mathrm{Nx} = 512, \mathrm{Ny} = 512$ 、時間步數為 $\mathrm{Nt} = 5000$ 。空間計算使用快速傅立葉轉換(FFT)方法、時間積分則採用四階 Runge-Kutta (RK4)方法。一共生成 14 筆數據，其中第 14 筆數據作為 nature run 使用，如 fig.1(a)所示。

C. U-Net model

U-Net 模型是一種基於機器學習的模型，將其套用於預測渦度場的未來狀態，該模型架構如附圖所示，目標是預測 100 個時間步後的渦度場。具體來說如 fig.2 所示 U-Net 模型以時間 $t = t_0$ 的渦度場作為輸入、輸出時間 $t = t_0 + 100$ 的渦度場，利用數值模式前 13 筆數據作為 U-Net 模型的訓練數據。相比於傳統的數值方法 U-Net 模型具有顯著的計算優勢，ML_v3 版本的 U-Net 模型每次計算僅需 0.2 秒，而傳統基於 FFT 的方法則需 300 秒完成相同的計算，如 fig.1(a)，這些效能提升使 U-Net 模型成為預報和分析渦度場的高效工具，更詳細系的解釋可以從我 GitHub 的 Readme_rotation.txt 得知。

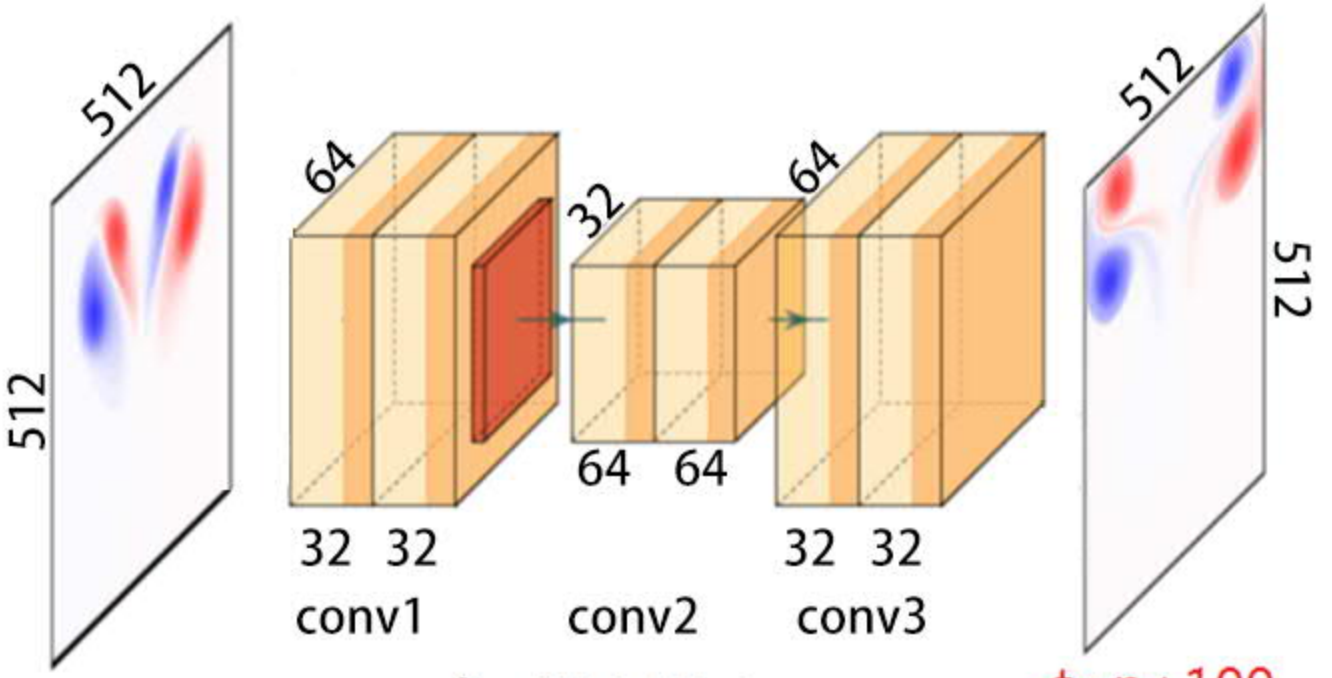

512
512
t=n
64
32
64
64
64
32
32
32
32
conv1
conv2
conv3
512
512
t=n+100

fig.2 U-Net

被訓練出來的 U-Net 模型將作為1.5 提到的 M 模式

3. Experimental design

在實驗中首先面臨的挑戰是處理 $(x,y)=(512,512)$ 的渦度場，計算 512^2 個變數導致矩陣運算的規模達到 $(512^2,512^2)$ ，如此龐大的運算需求對計算資源構成極大挑戰，為了解決此問題我將模式配合實驗，把實驗的網格內插至模式所需的網格點，計算完後將其 regrid 回 $(x,y)=(32,32)$ 的實驗網格大小。

我想測試多種參數設定以評估其對結果的影響，並挑選出表現最佳的參數組合。以下是測試的主要變數：Localization 的 L 、Inflation 的 $inflation\_alpha$、ensemble member 數量 K ，以及 regrid 網格大小 $(regridsize, regridsize)$ 。

A. Localization

EnKF 通常需要對從系集中計算的誤差協方差進行局地化，這是因為觀測影響需要限制在一定範圍內，以應對以下問題：

- 採樣誤差(sampling errors)
- 分析自由度的限制
- 計算成本的考量
- 本次實驗必須 localize，因為迭帶 3 次以上就會爆掉 fig.3(a)

我採用了 Greybush(2011)提出的 B-localization

$$P_{loc}^{f}=\left[\exp\left(-\frac{d(i',j')^2}{2L^2}\right)\right]_{ij} P^f \quad \cdots(3)$$

其中 d 為格點 (i,j) 與 $(i'j')$ 的距離， L 為局地化範圍參數。但是這樣會有一點問題，詳細在 fig.3(b)有說明在 2D 情況下這個公式會出現一點問題，但目前我就先套用方程式(3)。

我測試了不同的局地化範圍 L ，分別為 2 、5 、10 、15 ，並固定 $inflation\ alpha\ =0$, $K=20$, $(x,y)=(32,32)$。

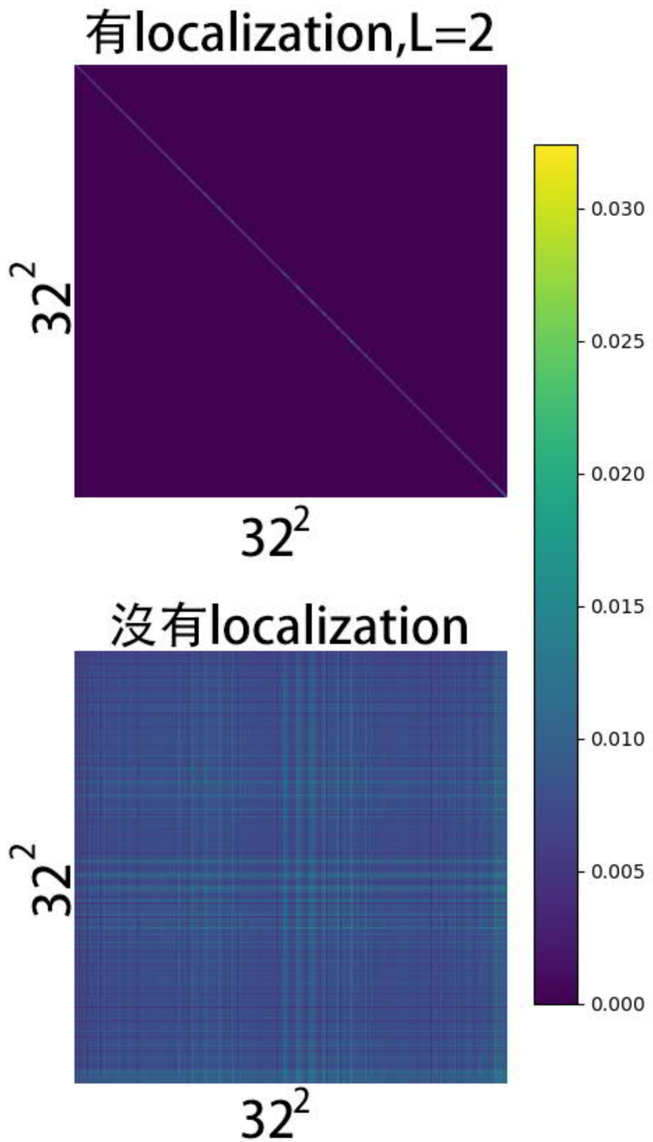
B. Inflation

在 EnKF 中 inflation 的目的是解決系集分散度不足的問題，即隨時間演進系集的變異性可能縮小，導致誤差協方差的低估。

fig.3 localization

(a)

t=3，預報3次後沒有localization明顯惡化



(b)

老師上課說的應該是1D的Localization，但是我這是2D的情況。

我會嘗試說明要怎麼做但是我應該不會這麼做 :3

先用(x,y)=(10,10)作為舉例:我們將其變成P圖應該會是(100,100)

而點A應該是第66個所以會在(66,66)

但是與A相鄰的a卻是在下面這些點:

(55,55),(56,56),(57,57)

(65,65),　　,(67,67)

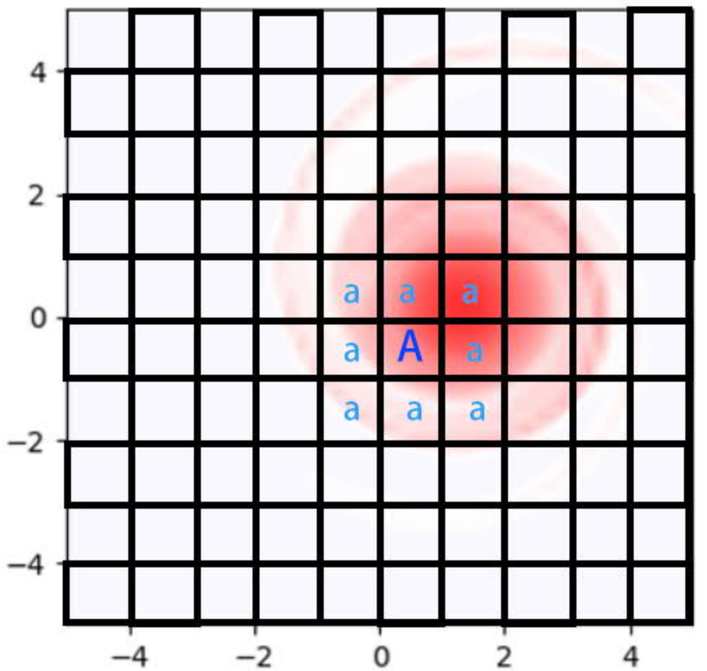
(75,75),(76,76),(77,77)

所以P應該在以下這些點有數值

(66,55),(66,56),(66,57)

(66,65),(66,66),(66,67)

(66,75),(66,76),(66,77)



也就是說當2D的情況L=2以上就會變得十分複雜

這解釋了兩點:

1.為什麼一開始會馬上(迭帶3次)爆開?

>因為localization沒有做好，用十字很快傳染出去。

2.為什麼L=2可能會比L=10、L=15要好?

>因為其實這個套用1D的公式L越大反而會抓進更多實際距離與自己很遠的。相反L小與自己更相關，更接近單位矩陣。

我測試了不同的 inflation 係數 $inflation\_alpha$，分別為 0、0.1、0.15、0.2、0.3，並固定$L = 2$, $K = 20$, $(x, y) = (32,32)$。

C. Ensemble member 數量

系集成員的數量 K 會直接影響誤差協方差的估算準確性和運算成本，增加成員數量可以提升估算的準確性，也會顯著增加計算負擔。

我測試了三種系集成員數量 K，分別為 20、30、40，並固定$L = 2$, $inflation\ alpha = 0$, $(x, y) = (32,32)$。

D. Regrid

為了降低計算量，這些實驗都會對渦度場進行了 regrid，將其網格大小從原始的 $(x, y) = (512,512)$ 降至較低解析度。
測試的 regrid 大小包括 $(32,32)$ 和 $(64,64)$，以找出在固定$L = 2$, $inflation\ alpha = 0$, $K = 40$ 計算效率與結果準確性之間的最佳平衡。

另一組(E、F)不是調整參數的實驗但是跟 DA 不是太相關所以我只描述蓋念，但是應該會是有趣的問題。

E. 資料同化對於預報的改善幅度

得到一筆觀測資料 $t = n$ 我們可以對它逕行預報，得到這些時間的模式資料 $t = n + 1$, $t = n + 2$, … , $t = n + 7$。隨後得到一筆觀測資料 $t = n + 1$ 我們也可以對它逕行預報，所以我們有 7 筆描述在 $t = 7$ 的模式資料，這些資料是怎麼收斂的是一個我也想做的東西，但是和 DA 就不知道怎麼連結。

F. 可預報度

有做 DA 沒做 DA 在可預報度有沒有顯著提升。

4. Results and discussion

完整結果如圖 4 所示，圖 4 分為以下四部分：

(a) Localization
(b) Inflation
(c) Ensemble member 數量
(d) Regrid

每部分圖表分為兩個子圖：

- 左圖（誤差時序圖）： X 軸為同化時序，Y 軸為均方根誤差（RMSE）。
- 右圖（特定時間錯誤度-誤差圖）： X 軸為 RMSE，Y 軸為 Normalized Index（主要描述觀測資訊的錯誤率）。

通常情況下，當 Normalized Index 接近 0（觀測較為準確）時，RMSE 應該很小，整個圖形會接近“く”型結構。本實驗選擇同化時序 $t = 10$ 和 $t = 40$ 作為參考，並以菱形與圓形標示。

Normalized Index 的計算公式如下：

$$\text{Normalized index}_{\text{i,(k)}} = \frac{\text{x}^{a}_{\text{i,(k)}} - x^{a}_{i,determine}}{\sigma_{\text{i,(k)}}} \quad \cdots (4)$$

A. Localization

結果顯示 $L = 2$ 和 $L = 5$ 的表現較好，但是從特定時間錯誤度-誤差圖可以觀察到即使在觀測誤差較大的情況下 $L = 2$ 仍能保持良好表現，因此我將選擇 $L = 2$ 作為後續分析的基準。

B. Inflation

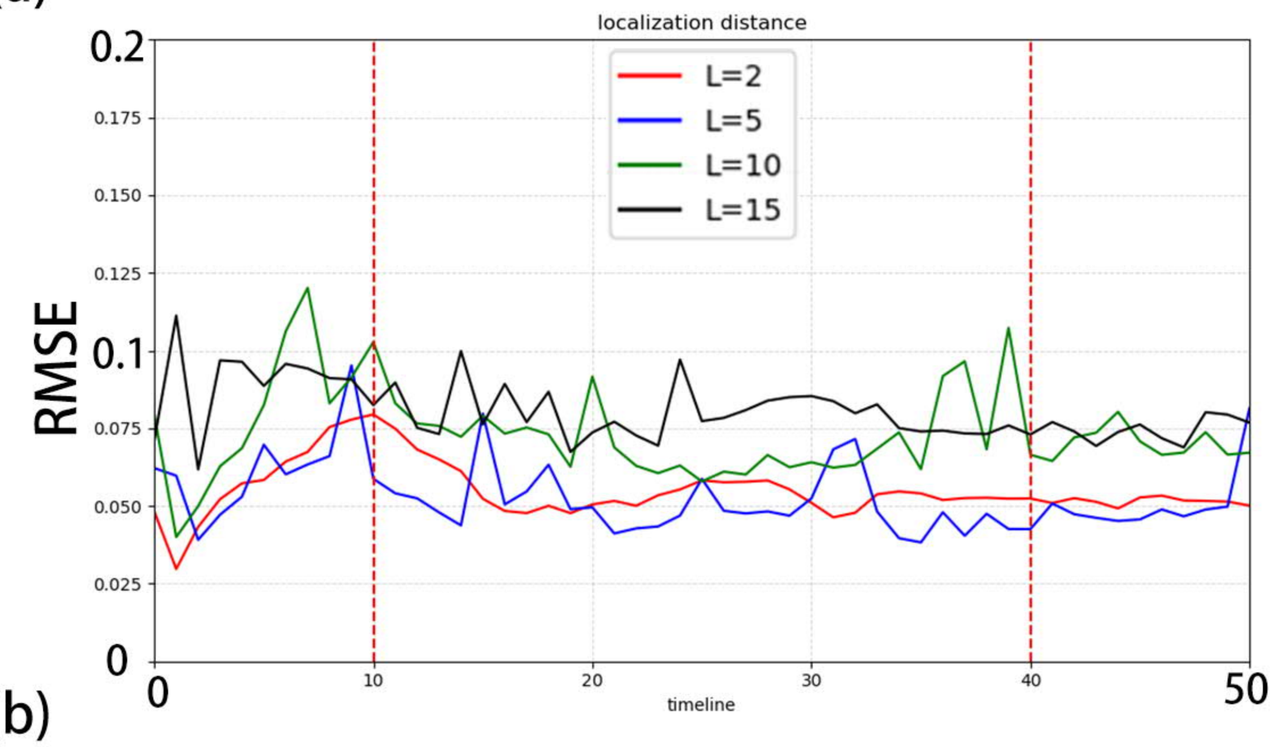
從誤差時序圖可以清楚看到 $inflation\ alpha\ = 0$ 的表現最佳。

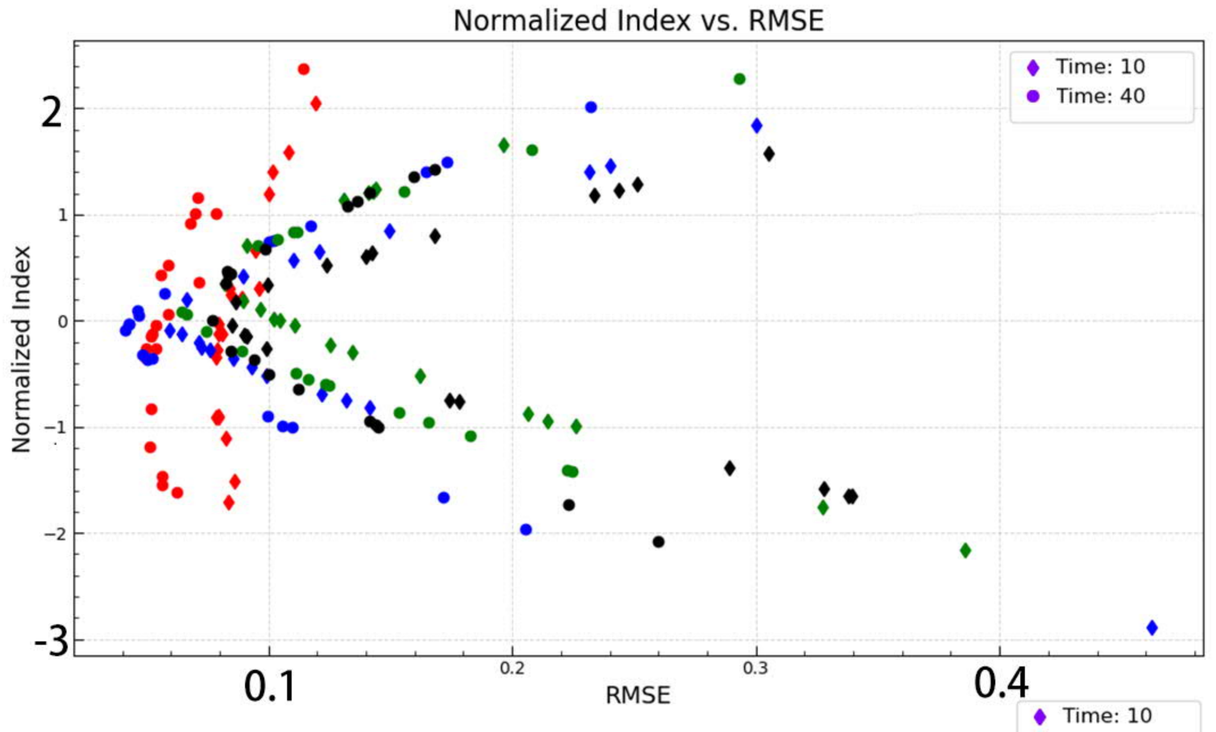
C. Ensemble member 數量

一般情況下當變數數量較多時，增加 K 通常會提升表現。但是在本案例中並未觀察到顯著的改善，不過根據特定時間錯誤度-誤差圖，我選擇 $K = 40$ 作為後續同化的基礎。

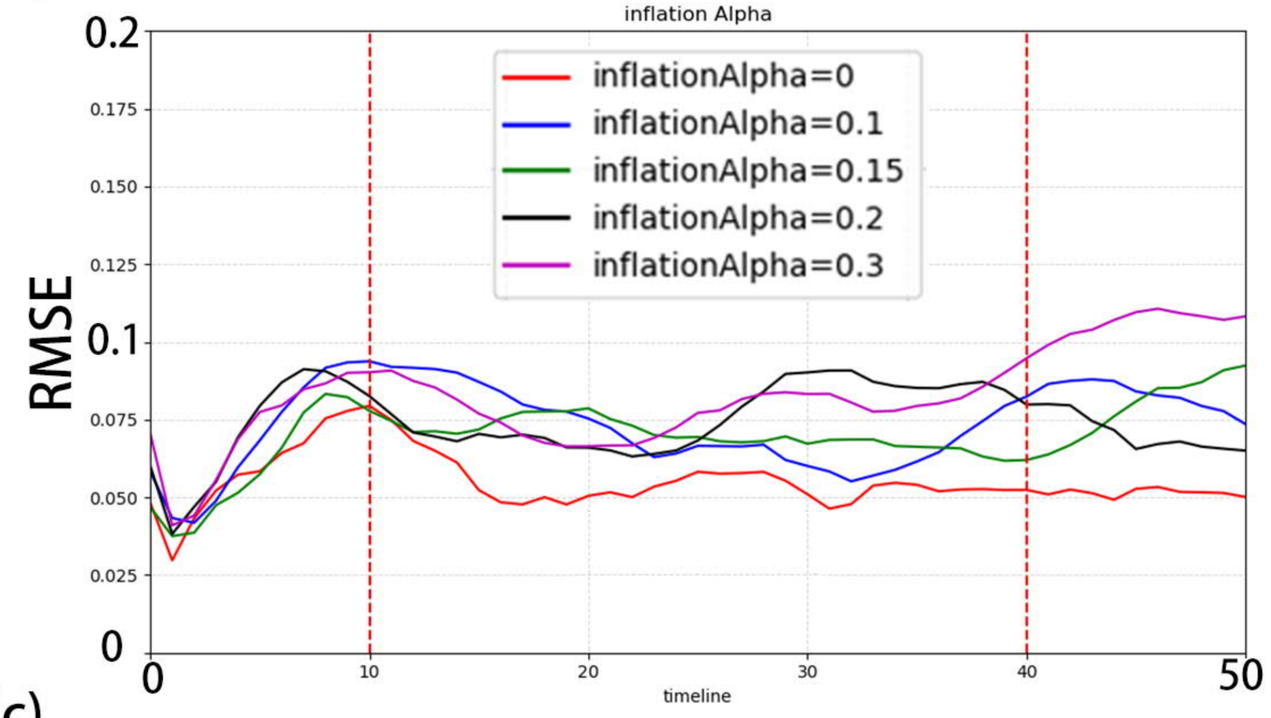
D. Regrid

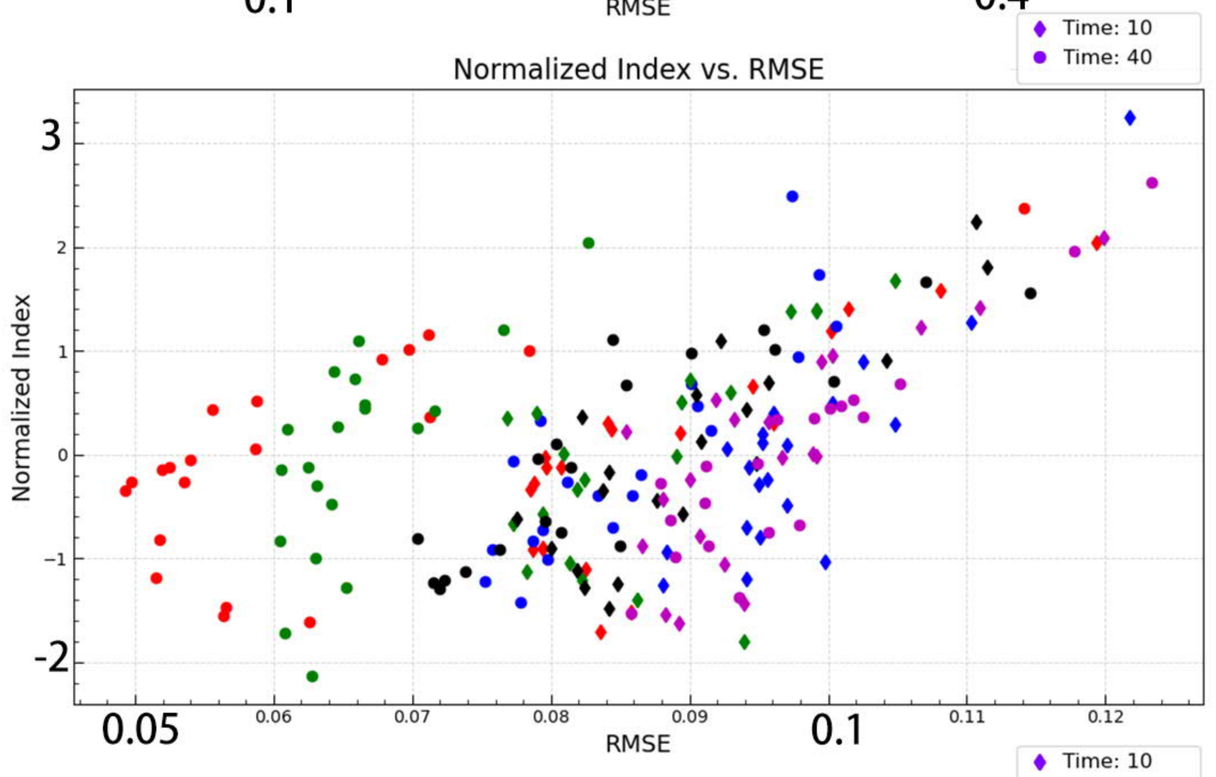
(a)



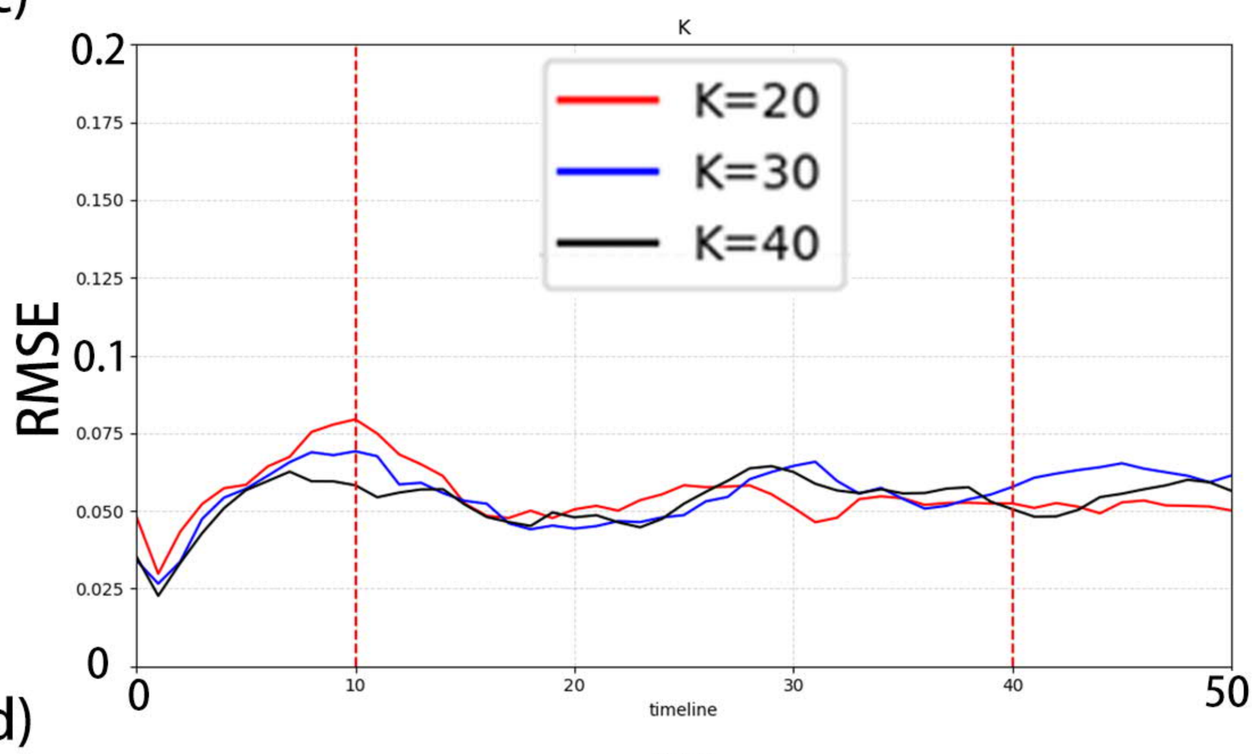


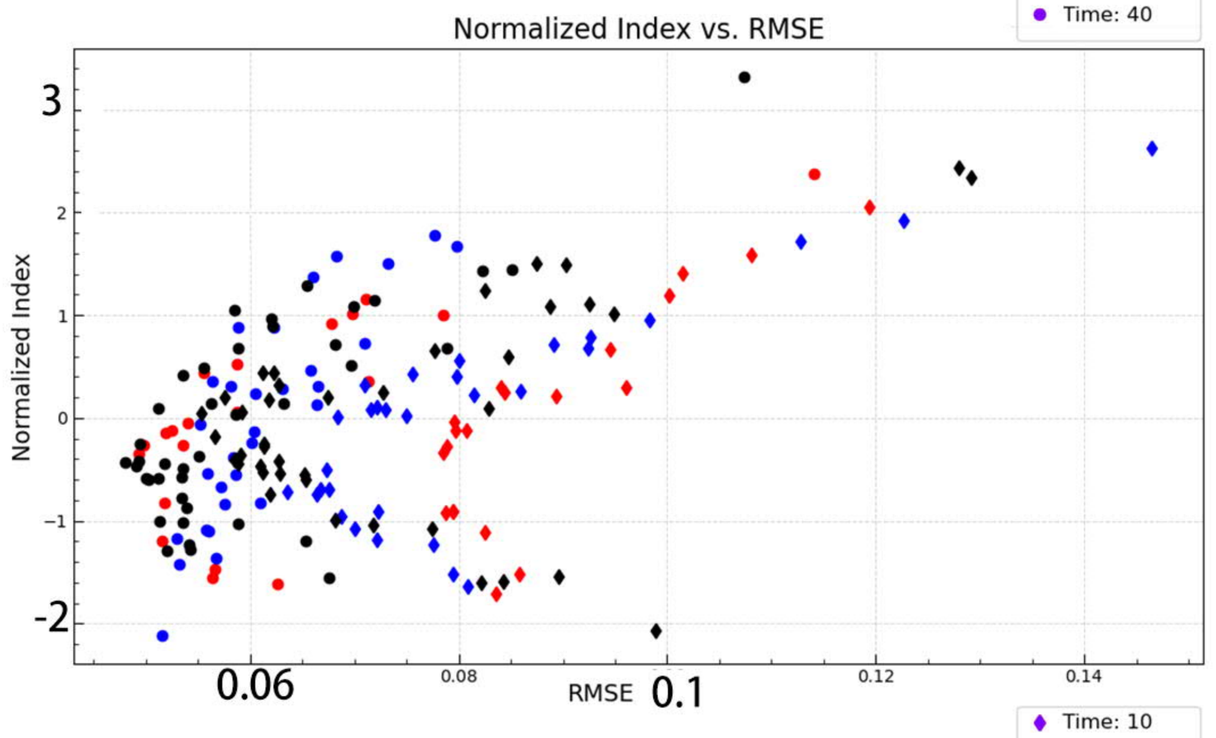
(b)



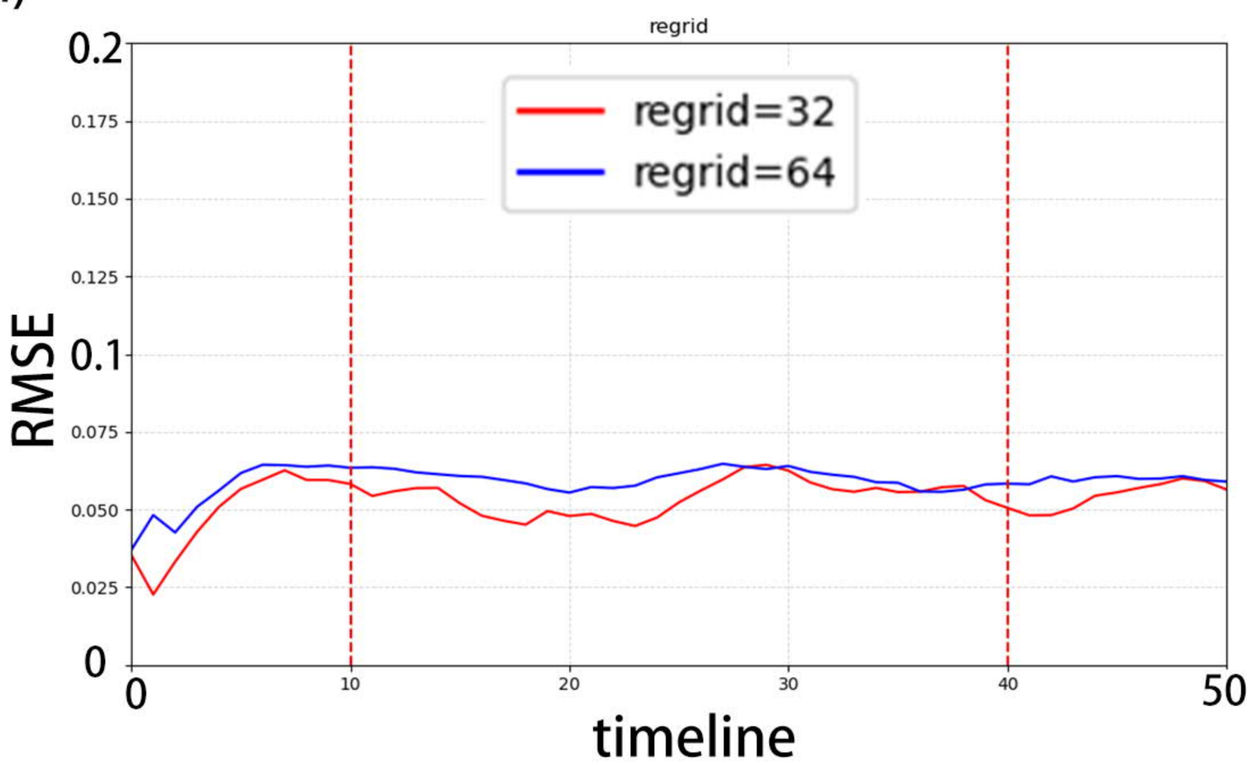


(c)





(d)



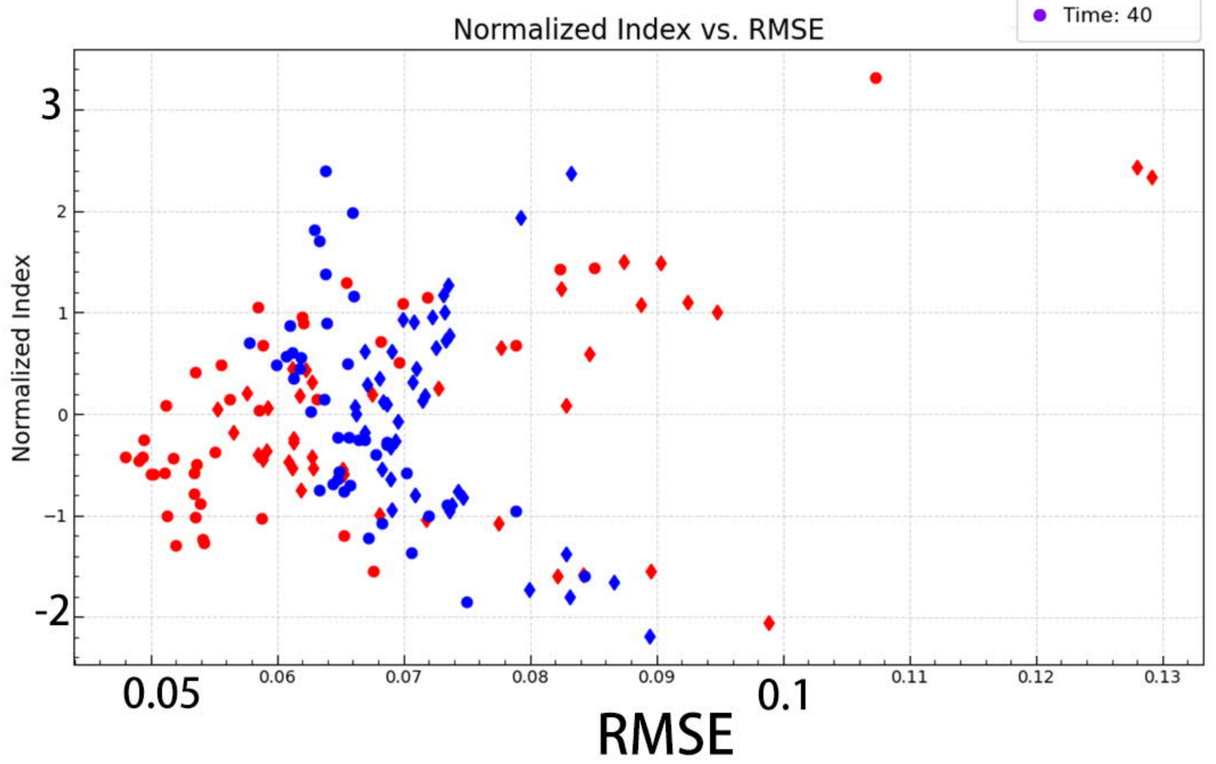


fig.4 result

從誤差時序圖與特定時間錯誤度-誤差圖可以清楚看到，regrid: $(x, y) = (32,32)$ 的網格化解析度表現較佳。

最後可以看一下同化的影片，我將其上傳到 Youtube: https://youtube.com/shorts/BpJTssM7YLg?feature=share

5. Conclusion

本研究通過對 Ensemble Kalman Filter（EnKF）資料同化系統的多參數測試，發現當局地化範圍 L 設為 2、inflation 係數設定為 0、系集成員數量 K 為 40，並選擇 regrid 解析度為(x, y) = (32, 32)時系統表現最佳。在誤差時序圖中，該組合顯示出最小的 RMSE，並且在特定時間的誤差度量中均能保持較低的誤差，顯示出良好的預測穩定性和準確性。

L=2 的局地化範圍有效控制了觀測數據的影響範圍，避免了過度調整，並提高了同化結果的可靠性，inflation alpha =0 則能夠保持較為真實的系集變異性，而 K=40 則在計算負擔與結果準確性之間取得了良好的平衡。這些結果表明，適當的參數設置對資料同化的效果至關重要，為後續的大氣預測和模型應用提供了重要的參考。

6. Programming code, provided in any of the following ways:

GitHub link: https://github.com/derek1403/2024DA-