

UNIVERSIDADE FEDERAL DE PELOTAS
Instituto de Física e Matemática
Programa de Pós-Graduação em Modelagem Matemática



Dissertação

Previsão por Conjuntos Agrupados por K-Means através do Sistema de Lorenz

Fernanda Batista da Silva

Pelotas, 2026

Fernanda Batista da Silva

Previsão por Conjuntos Agrupados por K-Means através do Sistema de Lorenz

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Matemática do Instituto de Física e Matemática da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Modelagem Matemática.

Orientador: Prof. Dr. Fabrício Pereira Harter
Coorientador: Prof. Dr. Leonardo Calvetti

Pelotas, 2026

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

S586p Silva, Fernanda Batista da

Previsão por conjuntos agrupados por K-Means através do sistema de Lorenz [recurso eletrônico] / Fernanda Batista da Silva ; Fabricio Pereira Harter, orientador ; Leonardo Calvetti, coorientador. — Pelotas, 2026.
101 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Modelagem Matemática, Instituto de Física e Matemática, Universidade Federal de Pelotas, 2026.

1. Previsão por conjuntos. 2. K-means. 3. Modelo de Lorenz. 4. Sistemas caóticos. I. Harter, Fabricio Pereira, orient. II. Calvetti, Leonardo, coorient. III. Título.

CDD 515.39

**PREVISÃO POR CONJUNTOS AGRUPADOS POR K-MEANS ATRAVÉS DO
SISTEMA DE LORENZ**

por

Fernanda Batista da Silva

Dissertação submetida ao Programa de Pós-Graduação em Modelagem Matemática, PPGMMat, do Instituto de Física e Matemática da Universidade Federal de Pelotas, como parte dos requisitos necessários para a obtenção do Título de

Mestre em Modelagem Matemática

Banca Examinadora:

Prof. Dr. Fabrício Pereira Harter – Orientador (UFPel)

Prof. Dr. Leonardo Calvetti – Co-Orientador (UFPel)

Prof. Dr. Mateus da Silva Teixeira (UFPel)

Profa. Dra. Julio Renato Quevedo Marques (UFPel)

Prof. Dr. Marcelo Romero de Moraes (Unipampa)

Pelotas, 08 de abril de 2026

AGRADECIMENTOS

Primeiramente, agradeço a Deus por ter sido o meu sustento, a minha força e o meu guia em todos os momentos, iluminando os meus passos e me dando discernimento quando os caminhos pareciam difíceis.

Agradeço imensamente aos meus filhos, Eduardo e Gustavo, que são a minha maior motivação e a razão de absolutamente tudo na minha vida. Obrigado pela paciência infinita nos momentos em que precisei me ausentar para estudar e por vibrarem comigo a cada pequena vitória. Este trabalho também é de vocês.

Olhando para trás, lembro-me do quanto foi difícil mudar para uma cidade nova, enfrentar o desconhecido e deixar tudo para trás em nome de um sonho e do desejo de estudar. A saudade e a distância foram desafios diários, mas que tornaram essa conquista ainda mais valorosa. À minha família, agradeço por ter sido o meu porto seguro e por segurarem a minha mão mesmo de longe, me dando o suporte necessário para que eu não desistisse.

Aos meus orientadores, Prof. Dr. Fabrício Pereira Harter e Prof. Dr. Leonardo Calvetti, pela condução brilhante, pela paciência e pelos ensinamentos valiosos que levarei para sempre na minha trajetória.

À coordenadora do programa, Prof^a Dr^a Daniela Buske, pelo acolhimento genuíno, pela dedicação extrema e por estar sempre pronta a estender a mão com um apoio fundamental em todos os momentos que precisei ao longo deste mestrado.

Aos professores da banca examinadora, Prof. Dr. Marcelo Romero de Moraes, Prof. Dr. Julio Renato Quevedo Marques e Prof. Dr. Mateus da Silva Teixeira, pelas ricas contribuições.

Ao Programa de Pós-Graduação em Modelagem Matemática da UFPel, aos professores e colegas, pelas trocas de experiências que tornaram essa jornada mais leve e enriquecedora, bem como à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio concedido para a realização deste trabalho.

Você nunca sabe que resultados virão da sua ação. Mas se você não fizer nada, não existirão resultados.

— MAHATMA GANDHI

RESUMO

SILVA, Fernanda Batista da. **Previsão por Conjuntos Agrupados por K-Means através do Sistema de Lorenz**. Orientador: Fabrício Pereira Harter. 2026. 101 f. Dissertação (Mestrado em Modelagem Matemática) – Instituto de Física e Matemática, Universidade Federal de Pelotas, Pelotas, 2026.

A previsão por conjuntos (*ensemble forecasting*) é uma abordagem amplamente empregada para representar incertezas em sistemas dinâmicos sensíveis às condições iniciais. Neste trabalho, investiga-se o uso do algoritmo *K-means*, com inicialização *K-means++*, no processamento de previsões por conjuntos geradas com o modelo de Lorenz, adotado como referência para o estudo do comportamento caótico. O objetivo consiste em verificar se o agrupamento, aplicado de forma independente a cada variável de estado do sistema, pode identificar trajetórias representativas com menor erro do que a média aritmética simples do *ensemble*. Para isso, foram gerados conjuntos com diferentes números de membros e diferentes magnitudes de perturbação nas condições iniciais, permitindo analisar como o tamanho do *ensemble* e o nível de dispersão influenciam o desempenho do método. O número de grupos foi definido com base no Coeficiente de Silhueta, e a avaliação dos resultados foi realizada por meio da Raiz do Erro Quadrático Médio (RMSE), calculada em relação à trajetória de referência. Os experimentos foram organizados em horizontes temporais de curto, médio e longo prazo, com o objetivo de examinar o comportamento do agrupamento em diferentes estágios da evolução do sistema. Os resultados mostram que o desempenho do método depende da configuração do *ensemble* e do horizonte de previsão, apresentando resultados mais favoráveis em cenários de dispersão intermediária, nos quais a média aritmética simples deixa de representar adequadamente a estrutura do conjunto sem que as trajetórias tenham se tornado excessivamente dispersas. Os dados obtidos indicam que, nos cenários analisados, o *K-means* pode representar uma alternativa útil para a organização e a análise estrutural de previsões por conjuntos em sistemas caóticos de baixa dimensão.

Palavras-chave: previsão por conjuntos; k-means; modelo de Lorenz; sistemas caóticos.

ABSTRACT

SILVA, Fernanda Batista da. **Ensemble Forecasting Clustered by K-Means through the Lorenz System**. Advisor: Fabrício Pereira Harter. 2026. 101 f. Dissertation (Masters in Mathematical Modeling) – Institute of Physics and Mathematics, Federal University of Pelotas, Pelotas, 2026.

Ensemble forecasting is a widely used approach for representing uncertainties in dynamical systems that are sensitive to initial conditions. This study investigates the use of the *K-means* algorithm, with *K-means++* initialization, in the processing of ensemble forecasts generated with the Lorenz model, adopted as a reference for the study of chaotic behavior. The objective is to examine whether clustering, applied independently to each individual system variable, can identify representative trajectories with lower error than the arithmetic mean of the *ensemble*. To this end, ensembles were generated with different numbers of members and different magnitudes of perturbation in the initial conditions, making it possible to analyze how ensemble size and dispersion level influence the performance of the method. The number of clusters was defined on the basis of the Silhouette Coefficient, and the results were evaluated using the Root Mean Square Error (RMSE), computed with respect to the reference trajectory. The experiments were organized into short-, medium-, and long-range forecast horizons in order to examine the behavior of clustering at different stages of system evolution. The results show that the performance of the method depends on the configuration of the *ensemble* and on the forecast horizon, with more favorable results in scenarios of intermediate dispersion, in which the arithmetic mean no longer adequately represents the structure of the ensemble without the trajectories having yet become excessively dispersed. The data obtained indicate that, in the scenarios analyzed, *K-means* may represent a useful alternative for the organization and structural analysis of ensemble forecasts in low-dimensional chaotic systems.

Keywords: ensemble forecasting; k-means; Lorenz model; chaotic systems.

LISTA DE FIGURAS

Figura 1	Representação tridimensional do atrator estranho de Lorenz no espaço de fases (X, Y, Z) para o regime caótico ($r = 32$), evidenciando a estrutura bimodal conhecida como asas da borboleta.	27
Figura 2	Divergência de trajetórias no modelo de Lorenz nos regimes (a) caótico ($r = 32$) e (b) não linear estável ($r = 10$). A simulação apresenta a evolução temporal das variáveis de estado (X, Y, Z) a partir da condição inicial de controle (linha sólida) e da condição perturbada em 10^{-1} (linha tracejada).	28
Figura 3	Fluxograma das etapas metodológicas adotadas neste estudo. . . .	36
Figura 4	Comportamento das variáveis X, Y e Z no curto prazo ($0 \leq t < 6$) para o cenário de 100 membros e ruído de 10^{-2} . As trajetórias dos membros apresentam alta convergência em relação à referência, evidenciando a baixa dispersão do <i>ensemble</i> nesse intervalo. . . .	46
Figura 5	Comportamento das variáveis X, Y e Z no médio prazo ($6 \leq t < 13$) para o cenário de 100 membros e ruído de 10^{-2} . O aumento da dispersão evidencia a limitação da média aritmética simples em representar a estrutura do <i>ensemble</i> nesse intervalo.	48
Figura 6	Comportamento das variáveis X, Y e Z no longo prazo ($13 \leq t < 20$) para o cenário de 100 membros e ruído de 10^{-4} . A maior dispersão entre as trajetórias templates evidencia a limitação da média aritmética simples em representar o comportamento do <i>ensemble</i> nesse intervalo.	49
Figura 7	Comportamento das variáveis X, Y e Z no curto prazo ($0 \leq t < 6$) para o cenário de 100 membros e perturbação de 10^{-2} , com aplicação do <i>K-means++</i> . As trajetórias apresentam alta convergência espacial, condição em que a média aritmética simples permanece estável e o agrupamento não apresenta vantagem nessa configuração específica.	52
Figura 8	Comportamento das variáveis X, Y e Z no médio prazo ($6 \leq t < 13$) para o cenário de 100 membros e ruído de 10^{-2} , com aplicação do <i>K-means++</i> . O gráfico mostra uma situação em que o agrupamento identifica uma trajetória mais próxima da referência do que a média aritmética simples.	54

Figura 9	Comportamento das variáveis X , Y e Z no longo prazo ($13 \leq t < 20$) para o cenário de 100 membros e perturbação de 10^{-2} , com aplicação do <i>K-means++</i> . O gráfico ilustra a saturação da dispersão em que a distribuição das trajetórias ocupa de forma homogênea os lóbulos do atrator estranho.	55
Figura 10	Distribuição espacial das variáveis X , Y e Z no curto prazo ($0 \leq t < 6$) para o cenário com 100 membros e ruído de 10^{-2} . A concentração dos membros no instante final do intervalo evidencia a baixa dispersão do conjunto e a proximidade entre a média aritmética simples e a trajetória de referência.	58
Figura 11	Distribuição espacial das variáveis X , Y e Z no médio prazo ($6 \leq t < 13$) para o cenário com 100 membros e ruído de 10^{-2} . A configuração no instante final mostra maior separação entre regiões do conjunto, favorecendo a proximidade entre o centróide selecionado e a trajetória de referência.	60
Figura 12	Distribuição espacial das variáveis X , Y e Z no longo prazo ($13 \leq t < 20$) para o cenário com 100 membros e ruído de 10^{-2} . A maior dispersão dos membros no instante final reduz o contraste entre a média aritmética simples e os centróides como representantes da trajetória de referência.	61
Figura 13	Distribuição espacial no médio prazo ($6 \leq t < 13$) para o cenário com 100 membros e ruído de 10^{-3} . A maior separação entre os membros favorece a identificação de grupos mais definidos.	65
Figura 14	Distribuição espacial no médio prazo ($6 \leq t < 13$) para o cenário com 15 membros e ruído de 10^{-2} . A menor quantidade de pontos reduz o detalhamento da distribuição do conjunto.	67
Figura 15	Distribuição espacial no médio prazo ($6 \leq t < 13$) para o cenário com 50 membros e ruído de 10^{-2} . O aumento do número de membros permite visualizar com maior clareza a organização do conjunto.	68
Figura 16	Curto prazo ($0 \leq t < 6$), $N = 15$ e $\sigma_\eta = 10^{-4}$	76
Figura 17	Curto prazo ($0 \leq t < 6$), $N = 50$ e $\sigma_\eta = 10^{-2}$	77
Figura 18	Curto prazo ($0 \leq t < 6$), $N = 100$ e $\sigma_\eta = 10^{-4}$	78
Figura 19	Dispersão no curto prazo ($0 \leq t < 6$), $N = 15$ e $\sigma_\eta = 10^{-4}$	79
Figura 20	Médio prazo ($6 \leq t < 13$), $N = 100$ e $\sigma_\eta = 10^{-4}$	81
Figura 21	Médio prazo ($6 \leq t < 13$), $N = 50$ e $\sigma_\eta = 10^{-4}$	82
Figura 22	Médio prazo ($6 \leq t < 13$), $N = 50$ e $\sigma_\eta = 10^{-3}$	83
Figura 23	Médio prazo ($6 \leq t < 13$), $N = 100$ e $\sigma_\eta = 10^{-4}$	84
Figura 24	Dispersão no médio prazo ($6 \leq t < 13$), $N = 15$ e $\sigma_\eta = 10^{-4}$	85
Figura 25	Dispersão no médio prazo ($6 \leq t < 13$), $N = 15$ e $\sigma_\eta = 10^{-3}$	86
Figura 26	Dispersão no médio prazo ($6 \leq t < 13$), $N = 50$ e $\sigma_\eta = 10^{-4}$	87
Figura 27	Longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-4}$	89
Figura 28	Longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-3}$	90
Figura 29	Longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-4}$	91
Figura 30	Longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-3}$	92
Figura 31	Longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-4}$	93
Figura 32	Longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-3}$	94
Figura 33	Dispersão no longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-4}$	95

Figura 34	Dispersão no longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-3}$	96
Figura 35	Dispersão no longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-4}$	97
Figura 36	Dispersão no longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-3}$	98
Figura 37	Dispersão no longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-4}$	99
Figura 38	Dispersão no longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-3}$	100

LISTA DE TABELAS

Tabela 1	Distribuição geral dos casos em que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples, por horizonte temporal.	50
Tabela 2	Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE no curto prazo ($0 \leq t < 6$), no cenário com ruído de 10^{-2}	52
Tabela 3	Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE no médio prazo ($6 \leq t < 13$), para 100 membros.	54
Tabela 4	Frequência de seleção de K por horizonte temporal.	56
Tabela 5	Distribuição dos casos em que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples, por variável.	62
Tabela 6	Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE por nível de ruído, para a variável X com 100 membros.	66
Tabela 7	Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE por número de membros, para a variável X com ruído de 10^{-2}	69
Tabela 8	Resultados do horizonte temporal de curto prazo ($0 \leq t < 6$).	80
Tabela 9	Resultados do horizonte temporal de médio prazo ($6 \leq t < 13$).	88
Tabela 10	Resultados do horizonte temporal de longo prazo ($13 \leq t < 20$).	101

LISTA DE ABREVIATURAS E SIGLAS

ECMWF	<i>European Centre for Medium-Range Weather Forecasts</i>
EnKF	Filtro de Kalman de Conjunto (<i>Ensemble Kalman Filter</i>)
MAE	Erro Médio Absoluto (<i>Mean Absolute Error</i>)
MM5	<i>Fifth-Generation Penn State/NCAR Mesoscale Model</i>
MSC	<i>Meteorological Service of Canada</i>
NCAR	<i>National Center for Atmospheric Research</i>
NCEP	<i>National Centers for Environmental Prediction</i>
NMC	<i>National Meteorological Center</i>
PNT	Previsão Numérica de Tempo
RK4	Runge-Kutta de quarta ordem
RMSE	Raiz do Erro Quadrático Médio (<i>Root Mean Square Error</i>)
VDA	Assimilação Variacional de Dados
WCSS	<i>Within-Cluster Sum of Squares</i>
WPC	<i>Weather Prediction Center</i>
WRF	<i>Weather Research and Forecasting Model</i>

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Objetivo Geral	17
1.2	Objetivos Específicos	17
1.3	Justificativa	18
2	FUNDAMENTAÇÃO TEÓRICA	19
2.1	Previsão por Conjuntos	19
2.2	O Modelo de Lorenz	24
2.2.1	Formulação Matemática e Parâmetros Físicos	24
2.2.2	Estabilidade e Análise dos Pontos Críticos	25
2.2.3	Divergência de Trajetórias e Regimes Dinâmicos	26
2.3	Aprendizado de Máquina Não Supervisionado	29
2.3.1	Algoritmo K-means e Critérios de Agrupamento	30
3	METODOLOGIA	35
3.1	Integração Numérica do Modelo de Lorenz	36
3.2	Geração do Ensemble	37
3.3	Definição dos Horizontes Temporais	38
3.4	Estrutura dos Dados para Agrupamento	39
3.5	Seleção do Número de Grupos	40
3.6	Aplicação do Algoritmo K-means++	41
3.7	Definição do Centroide de Menor RMSE	42
3.8	Estimadores Comparados	43
3.9	Métricas de Avaliação	43
3.10	Análise Geométrica Complementar da Dispersão	44
4	RESULTADOS E DISCUSSÕES	45
4.1	Análise do <i>Ensemble</i> referente à Média Aritmética Simples	45
4.1.1	Convergência e Sobreposição no Curto Prazo	45
4.1.2	Divergência e a limitação da Média no Médio Prazo	46
4.1.3	Saturação da Dispersão no Longo Prazo	48
4.2	Análise pelos Horizontes Temporais utilizando <i>K-means++</i>	50
4.2.1	Curto Prazo ($0 \leq t < 6$)	50
4.2.2	Médio Prazo ($6 \leq t < 13$)	52
4.2.3	Longo Prazo ($13 \leq t < 20$)	54
4.3	Seleção do Número de Agrupamentos (K)	55
4.4	Análise da Distribuição Espacial dos Agrupamentos	57
4.4.1	Distribuição Espacial no Curto Prazo	57

4.4.2	Distribuição Espacial no Médio Prazo	58
4.4.3	Distribuição Espacial no Longo Prazo	60
4.5	Análise Comparativa das Variáveis X, Y e Z	61
4.5.1	Análise das Variáveis X e Y	62
4.5.2	Análise da Variável Z	63
4.5.3	Comparação entre as Variáveis ao Longo do Tempo	63
4.6	Análise Comparativa dos Níveis de Ruído	64
4.6.1	Sensibilidade à Densidade do Ensemble	66
5	CONCLUSÕES	70
	REFERÊNCIAS	72
	APÊNDICES	75
	APÊNDICE A – ANÁLISE GRÁFICA E TABELAS COMPLEMENTARES DO HORIZONTE TEMPORAL DE CURTO PRAZO	76
	APÊNDICE B – ANÁLISE GRÁFICA E TABELAS COMPLEMENTARES DO HORIZONTE TEMPORAL DE MÉDIO PRAZO	81
	APÊNDICE C – ANÁLISE GRÁFICA E TABELAS COMPLEMENTARES DO HORIZONTE TEMPORAL DE LONGO PRAZO	89

1 INTRODUÇÃO

A melhoria da precisão da previsão do tempo constitui um desafio contínuo. As equações que descrevem os fenômenos da atmosfera são não lineares e apresentam sensibilidade extrema às condições iniciais, de modo que pequenas variações podem levar a trajetórias significativamente divergentes ao longo do tempo de integração dos modelos de previsão (LORENZ, 1963). Essas incertezas, associadas também a limitações como a resolução finita dos modelos numéricos, as parametrizações de processos físicos e as imprecisões nas medições iniciais, reduzem a confiabilidade das previsões, sobretudo em horizontes temporais mais longos (KALNAY, 2003).

A previsão por conjuntos busca enfrentar essas limitações por meio da geração de múltiplas simulações com condições iniciais ou parâmetros perturbados (LEUTBECHER; PALMER, 2008). A variabilidade entre os membros do conjunto reflete o grau de incerteza associado à previsão: menor variabilidade indica maior previsibilidade, enquanto a média do conjunto tende a ser mais confiável do que uma previsão determinística isolada (KALNAY, 2003). Adicionalmente, estudos contemporâneos reafirmam que o mapeamento dessas dispersões em sistemas não lineares continua sendo a base para o aprimoramento dos modelos preditivos modernos (BAUER et al., 2015).

Inspirada na previsão de Monte Carlo, que utiliza perturbações aleatórias para representar incertezas (LEITH, 1974), esta dissertação aplica a previsão por conjuntos ao modelo de Lorenz, um sistema dinâmico de baixa dimensão capaz de reproduzir comportamento caótico análogo, em termos conceituais, ao da atmosfera (LORENZ, 1963). Reconhecido como um *toy model*, o modelo de Lorenz é particularmente adequado para o estudo de incertezas em sistemas sensíveis às condições iniciais, pois permite investigar essas questões em um ambiente matemático controlado e de menor complexidade (KALNAY, 2003).

Para lidar com a divergência e o espalhamento das trajetórias geradas, esta pesquisa emprega uma técnica de agrupamento não supervisionado via *K-means* com inicialização *K-means++*. O método é utilizado para organizar os membros do conjunto em grupos, empregando o Coeficiente de Silhueta para a definição dinâmica do

número de grupos (K). O algoritmo busca identificar padrões dinâmicos e regimes de comportamento que a média aritmética simples pode não representar adequadamente, permitindo analisar a previsibilidade em diferentes estágios do regime dinâmico (ARTHUR; VASSILVITSKII, 2007). Nessa perspectiva, investiga-se de que maneira o número de membros do *ensemble* influencia a redução dos erros e a obtenção de trajetórias representativas em um sistema dinâmico caótico de baixa dimensão (WILKS, 2011).

Esta dissertação está estruturada em cinco capítulos. No Capítulo 2, apresenta-se o referencial teórico, abordando os fundamentos da previsão por conjuntos, o modelo de Lorenz e o algoritmo *K-means*. No Capítulo 3, descreve-se a metodologia, incluindo a discretização do modelo, a geração do conjunto sob diferentes magnitudes de perturbação nas condições iniciais (ruído) e as etapas do agrupamento segmentado por horizontes temporais. No Capítulo 4, são analisados os resultados, com ênfase na qualidade dos grupos via silhueta e na redução relativa do RMSE em relação à trajetória de referência. Por fim, no Capítulo 5, apresentam-se as conclusões do trabalho e discutem-se suas limitações.

1.1 Objetivo Geral

Avaliar a previsão por conjuntos por meio de aprendizado de máquina não supervisionado, com base no algoritmo *K-means*.

1.2 Objetivos Específicos

- Implementar o modelo caótico de Lorenz para gerar previsões por conjuntos, considerando distintas configurações de tamanho de *ensemble* e magnitudes de perturbação inicial;
- Aplicar o algoritmo *K-means++* a cada uma das variáveis dinâmicas de forma isolada, para identificar e agrupar trajetórias com evoluções semelhantes no espaço de fases;
- Determinar o número ideal de agrupamentos em cada cenário operacional a partir do critério estatístico do Coeficiente de Silhueta;
- Comparar o desempenho preditivo dos centroides gerados pelo agrupamento em relação à média aritmética simples do *ensemble*, utilizando a métrica do RMSE;
- Avaliar o comportamento e a estabilidade das estruturas agrupadas ao longo de horizontes temporais de curto, médio e longo prazo.

1.3 Justificativa

A compreensão do escoamento atmosférico é fundamental para diversas atividades humanas, o que justifica a busca por diferentes estratégias de previsão. Entre essas estratégias, destacam-se os modelos numéricos de previsão de tempo, compostos por sistemas de equações diferenciais altamente dependentes das condições iniciais. Nesse contexto, torna-se relevante investigar abordagens que permitam compreender e representar, de forma mais informativa, a incerteza associada às soluções desses sistemas, como ocorre na previsão por conjuntos (PALMER, 2006).

Os chamados *toy models* são particularmente úteis para testar novas metodologias, pois reproduzem características dinâmicas de sistemas mais complexos com baixo custo computacional e maior facilidade de interpretação (LORENZ, 1963; KALNAY, 2003). Nesse sentido, o modelo de Lorenz constitui um ambiente matemático controlado e conceitualmente adequado para uma investigação exploratória, permitindo analisar, em baixa dimensão, o comportamento de trajetórias sensíveis às condições iniciais sem a complexidade adicional presente em modelos atmosféricos operacionais. A utilização de aprendizado de máquina não supervisionado, especificamente do algoritmo *K-means++*, fundamenta-se em sua capacidade de organizar dados em grupos com base na proximidade estatística, identificando padrões que a média simples do conjunto pode não representar de forma satisfatória (ARTHUR; VASSILVITSKII, 2007).

A presente investigação utiliza a técnica de agrupamento de forma univariada nas componentes do sistema. Partindo da abordagem de agrupamento de previsões já discutida na literatura (ALHAMED et al., 2002), este estudo analisa como essa técnica se comporta ao tratar as variáveis de forma independente sob diferentes níveis de ruído e diferentes dimensões de conjunto. Essa escolha se justifica pela necessidade de verificar se o algoritmo mantém sua capacidade de identificar trajetórias representativas quando aplicado individualmente às variáveis X , Y e Z , evitando que a maior amplitude de oscilação de uma componente mascare padrões dinâmicos relevantes das demais (WILKS, 2011).

Dessa forma, a pesquisa busca contribuir para a representação e interpretação da incerteza inerente à previsão de tempo por conjuntos, propondo um método alternativo de agrupamento estatístico para organizar suas trajetórias. Além de explorar a viabilidade dessa abordagem em um sistema caótico simplificado, o estudo também pode servir de referência metodológica para investigações futuras em modelos atmosféricos mais complexos, com potencial de ampliar a interpretabilidade dos dados e contribuir para o aprimoramento da qualidade das previsões numéricas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Previsão por Conjuntos

A previsão numérica do tempo (PNT) utiliza modelos computacionais para simular a evolução da atmosfera por meio da resolução de equações diferenciais que descrevem processos físicos e químicos, tais como temperatura, pressão e vento (KALNAY, 2003). Entretanto, a atmosfera se comporta como um sistema dinâmico não linear e altamente sensível às condições iniciais, de modo que pequenas incertezas no estado inicial podem ser amplificadas ao longo do tempo, o que limita a precisão das previsões determinísticas. Nesse contexto, a previsão por conjuntos busca representar essa limitação de forma probabilística por meio da geração de múltiplas simulações com variações controladas nas condições iniciais e, em alguns casos, nas parametrizações físicas, permitindo descrever a evolução futura em termos probabilísticos, e não por uma única trajetória determinística (LEUTBECHER; PALMER, 2008).

Epstein (1969) lançou as bases da previsão por conjuntos ao formalizar a abordagem estocástico-dinâmica. Nesse contexto, as variáveis atmosféricas passam a ser tratadas não como valores fixos, mas como distribuições de probabilidade, reconhecendo-se a existência de incerteza inerente ao estado inicial da atmosfera. Para tornar essa ideia aplicável, o autor desenvolveu uma formulação baseada na representação estocástica de múltiplos estados possíveis do sistema, estabelecendo a noção de que a previsão atmosférica pode ser tratada em termos probabilísticos, e não apenas determinísticos.

Inspirado por esses avanços, Leith (1974) investigou a habilidade teórica de previsões de *Monte Carlo* como aproximação da técnica de previsão estocástico-dinâmica proposta por Epstein. O procedimento consiste em introduzir perturbações aleatórias nas condições iniciais, simulando incertezas associadas à observação e à análise. Para testar a eficiência estatística dessa metodologia, Leith utilizou um modelo simplificado de turbulência bidimensional. Seus resultados mostraram que *ensembles* compostos por poucos membros já eram capazes de representar parte da incerteza associada à previsão, indicando a viabilidade prática da abordagem no contexto das

limitações computacionais então existentes.

Toth e Kalnay (1993) representaram um avanço importante ao implementarem o método de *breeding of growing modes* no *National Meteorological Center* (NMC). Essa metodologia baseia-se em um ciclo iterativo que busca isolar e amplificar os erros de análise que apresentam maior crescimento dentro do próprio modelo de previsão, simulando a forma como pequenas incertezas evoluem naturalmente na dinâmica atmosférica. Os autores mostraram que esse procedimento melhora a representação da variabilidade dos estados futuros e constitui uma estratégia dinâmica para a geração de perturbações iniciais. Paralelamente, Buizza e Palmer (1995) introduziram o uso de vetores singulares no *European Centre for Medium-Range Weather Forecasts* (ECMWF) para gerar perturbações otimizadas. Diferentemente do método anterior, os vetores singulares procuram identificar, no espaço de fase, as direções associadas ao maior crescimento do erro em um intervalo de tempo específico, consolidando o uso de aproximações lineares locais para a quantificação de incertezas em sistemas atmosféricos não lineares. Esses estudos indicam que a geração de perturbações iniciais evoluiu de abordagens empíricas para métodos dinamicamente orientados, capazes de representar de forma mais consistente o crescimento do erro (TOTH; KALNAY, 1993; BUIZZA; PALMER, 1995).

Kalnay *et al.* (1996) lideraram o projeto de Reanálise entre o *National Centers for Environmental Prediction* e o *National Center for Atmospheric Research* (NCEP/NCAR), utilizando um sistema de assimilação de dados mantido fixo ao longo do período de reanálise para criar um banco histórico consistente. A assimilação de dados consiste no processo matemático de combinar observações reais com os resultados do modelo numérico, a fim de obter a melhor estimativa possível do estado da atmosfera. Esse trabalho foi fundamental para definir uma base observacional coerente, amplamente utilizada na avaliação da qualidade e da dispersão de sistemas de previsão. Em continuidade a essa evolução, Krishnamurti *et al.* (1999) investigaram o *superensemble* multimodelo, uma abordagem que combina diferentes modelos globais por meio de regressão estatística ponderada. Os resultados evidenciaram que essa combinação demonstrou capacidade de mitigar vieses sistemáticos individuais através da atribuição de pesos diferenciados a cada componente, reforçando a premissa de que a diversidade estrutural entre formulações dinâmicas distintas pode contribuir para o ajuste estatístico e o estreitamento dos intervalos de erro do conjunto.

No início dos anos 2000, Stensrud, Bao e Warner (2000) investigaram como o número de membros e a representação de processos físicos afetam a previsão em mesoescala, que compreende fenômenos regionais de menor duração e maior intensidade. Para isso, utilizaram o Modelo de Mesoescala de 5ª Geração (MM5), sistema numérico regional desenvolvido pela *Pennsylvania State University* em conjunto com o NCAR para simular eventos atmosféricos de alta resolução. Os autores testaram

variações nas condições iniciais e nas parametrizações físicas, que são formulações matemáticas utilizadas para representar processos que ocorrem em escalas menores do que a resolução da grade do modelo. Entre essas parametrizações, destacam-se a microfísica, relacionada à formação de nuvens e precipitação, e a convecção, associada aos movimentos verticais de calor e umidade. Os resultados evidenciaram que a incerteza associada à física do modelo desempenha papel central na qualidade da previsão, mostrando que a confiabilidade do sistema depende do equilíbrio entre a quantidade de membros e a fidelidade física dos cenários simulados.

Com o aumento do volume de dados gerados pelos sistemas de previsão, Alhamed, Lakshmivarahan e Stensrud (2002) aplicaram análise de agrupamento (*cluster analysis*) em dados de *ensembles* multimodelo. Essa técnica consiste em organizar um conjunto de dados em grupos (*clusters*), de forma que elementos pertencentes ao mesmo grupo sejam mais semelhantes entre si do que em relação aos elementos de outros grupos. Para isso, os autores utilizaram o algoritmo *K-means*, método iterativo que particiona as simulações em K grupos distintos ao minimizar a distância entre cada membro e o centróide de seu respectivo grupo. A investigação revelou que os membros tendiam a se agrupar prioritariamente de acordo com o modelo de origem, e não apenas em função das variações nas condições iniciais. Esse resultado indicou que diferenças nas parametrizações físicas dos modelos podem exercer impacto superior ao das incertezas iniciais na estrutura final dos cenários, evidenciando a utilidade do agrupamento para diagnosticar a organização interna de sistemas de previsão.

Kalnay (2003) consolidou esses avanços em sua obra, destacando como a integração entre técnicas de assimilação de dados e métodos probabilísticos elevou a capacidade de representar a incerteza associada às previsões atmosféricas. Nesse contexto, o Filtro de Kalman por Conjuntos (*Ensemble Kalman Filter* – EnKF) é apresentado como um método estatístico que utiliza o próprio conjunto de previsões para estimar a incerteza do estado atual da atmosfera, permitindo um ajuste mais consistente entre o modelo e as observações. O autor reforça que sistemas de dinâmica não linear, nos quais pequenas variações iniciais podem gerar mudanças significativas no estado futuro, exigem abordagens probabilísticas. Essa perspectiva consolidou a necessidade de representar as incertezas inerentes ao caos por meio de múltiplas simulações, em vez de restringir a previsão a uma única trajetória determinística. Em conjunto, esses trabalhos mostram que o avanço da previsão por conjuntos depende tanto do refinamento das condições iniciais quanto da capacidade de interpretar estatisticamente a dispersão gerada pelo sistema.

Complementando os estudos anteriores, Yussouf, Stensrud e Lakshmivarahan (2004) realizaram uma análise de agrupamento em dados de previsão sobre a região da Nova Inglaterra. O estudo utilizou essa técnica para interpretar a estrutura

da dispersão (*spread*) do conjunto, isto é, o grau de incerteza associado à separação entre as diferentes previsões dos membros. A conclusão da pesquisa foi que a diversidade na configuração física dos modelos desempenha papel dominante na criação dessa dispersão. Isso sugere que o agrupamento constitui uma ferramenta eficaz para identificar a sensibilidade do sistema a diferentes processos físicos, auxiliando na organização de cenários que, de outra forma, poderiam parecer excessivamente dispersos e de difícil interpretação.

Em 2005, Wilks (2005) aprofundou o rigor estatístico ao detalhar métodos de verificação, como o Índice de Brier, métrica que avalia a conformidade de previsões probabilísticas por meio da diferença média quadrática entre a probabilidade prevista e o resultado observado. A decomposição desse índice permite separar a precisão do sistema de sua confiabilidade, evidenciando que uma previsão pode apresentar baixo erro e, ainda assim, carecer de calibração estatística, processo que consiste no ajuste corretivo das probabilidades previstas para que correspondam de forma fidedigna às frequências reais de ocorrência observadas historicamente. No mesmo ano, Buizza *et al.* (2005) realizaram um estudo comparativo entre os systems do ECMWF, MSC e NCEP, avaliando a capacidade de cada centro em manter uma dispersão suficiente para abranger os erros reais da atmosfera. Os resultados indicaram que, embora cada sistema apresentasse pontos fortes em diferentes horizontes de previsão, todos ainda enfrentavam o desafio de subestimar a incerteza em determinadas situações de transição atmosférica, resultando em conjuntos subdispersivos, nos quais a dispersão é inferior ao erro observado (WILKS, 2005; BUIZZA *et al.*, 2005).

Em 2008, Leutbecher e Palmer (2008) discutiram a introdução de parametrizações estocásticas no modelo global do ECMWF. Essa metodologia consiste em incorporar variações aleatórias nas parametrizações ou tendências associadas a processos físicos de pequena escala, denominados processos de subgrade, em vez de utilizar valores fixos. Os autores mostraram que essa abordagem pode elevar a dispersão do conjunto de forma mais realista e melhorar a confiabilidade das previsões probabilísticas. Em 2009, Johnson e Bowler (2009) investigaram a necessidade de calibração estatística para corrigir desvios estruturais e sistemáticos dos modelos, assegurando que o intervalo probabilístico estimado guarde correspondência matemática direta com a variabilidade observada. Para isso, utilizaram diagramas de confiabilidade, ferramenta gráfica que relaciona a probabilidade prevista com a frequência observada dos eventos. O estudo demonstrou que a aplicação de técnicas de calibração aproxima as previsões das frequências observadas, reduzindo vieses que poderiam induzir o tomador de decisão ao erro (LEUTBECHER; PALMER, 2008; JOHNSON; SWINBANK, 2009).

No cenário nacional, Calvetti (2011) aplicou o sistema de previsão por conjuntos em escala regional, utilizando o modelo *Weather Research and Forecasting* (WRF)

para prever eventos extremos de cheia na Bacia do Alto Iguaçu. A metodologia empregada envolveu a execução de múltiplas simulações que integraram variações nas condições iniciais e diferentes configurações de física do modelo. Os resultados demonstraram que a média do conjunto apresentou desempenho superior ao das previsões determinísticas tradicionais na detecção de picos de inundação, evidenciando o potencial da técnica para o aprimoramento de sistemas de alerta hidrológico no Brasil (CALVETTI; SILVA DIAS, 2011).

Em estudo recente, Lamberson *et al.* (2023) descreveram como o *Weather Prediction Center* (WPC) passou a organizar grandes volumes de dados em previsões de médio prazo. O estudo detalha como a aplicação do algoritmo *K-means* permite processar e sintetizar conjuntos de aproximadamente 100 simulações em um número reduzido de núcleos representativos. Esse processo viabiliza uma interpretação mais ágil das trajetórias mais prováveis e das alternativas plausíveis para os dias subsequentes. Os resultados evidenciaram que essa organização otimiza a comunicação da incerteza, permitindo que meteorologistas identifiquem tanto o cenário de maior consenso quanto cenários alternativos plausíveis que poderiam ser negligenciados em uma análise baseada apenas na média simples.

Em outra aplicação, Kiel e Colle (2024) compararam diferentes algoritmos de agrupamento aplicados ao estudo de ciclones extratropicais. Os autores demonstraram que a eficácia do agrupamento é fortemente determinada pelos critérios de semelhança adotados, e não apenas pelo algoritmo em si. O estudo concluiu que métricas que consideram o deslocamento e a morfologia das trajetórias produzem grupos mais coerentes e fisicamente interpretáveis. Essa abordagem permite uma caracterização mais robusta dos diferentes cenários de evolução desses sistemas, reduzindo o risco de que variações naturais de posicionamento prejudiquem a organização dos membros e conduzam a interpretações equivocadas sobre a intensidade do fenômeno.

Também em 2024, Jennrich *et al.* (2024) aplicaram o *K-means* para categorizar padrões de circulação de larga escala, denominados regimes de circulação. A investigação demonstrou que a previsibilidade atmosférica está relacionada à geometria do conjunto, tornando-se mais favorável quando as simulações convergem para grupos espaciais bem definidos. Para a presente pesquisa, essa observação é relevante, pois reforça a premissa de que a organização estrutural dos membros pode fornecer um diagnóstico mais informativo da dinâmica do sistema do que uma análise baseada exclusivamente na média simples.

Ao longo dessa trajetória, observa-se que a previsão por conjuntos evoluiu de uma formulação estocástico-dinâmica para um conjunto de técnicas voltadas não apenas à quantificação da incerteza, mas também à interpretação da estrutura interna dos cenários simulados. Essa perspectiva é particularmente importante no presente trabalho, no qual o interesse recai sobre a organização estrutural de trajetórias em um

sistema dinâmico caótico.

2.2 O Modelo de Lorenz

Entre os sistemas dinâmicos não lineares mais estudados na literatura, destaca-se o sistema de Lorenz, formado por três equações diferenciais ordinárias que se tornaram um marco na teoria do caos e no estudo da previsibilidade. Formulado na década de 1960 no contexto da modelagem simplificada da convecção atmosférica, esse sistema evidenciou que mesmo modelos determinísticos de baixa dimensão podem exibir comportamento aperiódico e sensibilidade extrema às condições iniciais Lorenz (1963). Pela relevância conceitual e pela riqueza de sua dinâmica, esse sistema passou a ocupar papel central na análise de regimes caóticos e dissipativos (GHIL; CHILDRESS, 1987).

A relevância desse modelo decorre do contraste entre a simplicidade de sua formulação matemática e a complexidade de sua dinâmica. Segundo Boyce, DiPrima e Meade (2024), o sistema serve como protótipo para o estudo de sistemas dissipativos que exibem comportamento caótico, refletindo propriedades essenciais de sistemas físicos reais nos quais a não linearidade impede a obtenção de soluções analíticas de longo prazo (GHIL; CHILDRESS, 1987).

Uma característica central do sistema é a sensibilidade extrema às condições iniciais. Lorenz (1963) observou que variações infinitesimais nos valores iniciais resultavam em trajetórias que divergiam exponencialmente após um curto intervalo de tempo. Essa descoberta estabeleceu que, em sistemas caóticos, a incerteza intrínseca à observação inicial impõe uma limitação prática à previsibilidade, uma vez que duas trajetórias que partem de estados arbitrariamente próximos podem divergir significativamente no espaço de fases.

2.2.1 Formulação Matemática e Parâmetros Físicos

As equações que governam a dinâmica do sistema estabelecem a relação entre as taxas de variação das variáveis de estado no tempo adimensional τ :

$$\frac{dX}{d\tau} = \sigma(Y - X) \quad (1)$$

$$\frac{dY}{d\tau} = rX - Y - XZ \quad (2)$$

$$\frac{dZ}{d\tau} = XY - \beta Z \quad (3)$$

Nessa formulação, as variáveis de estado descrevem propriedades associadas à dinâmica de uma célula de convecção fluida: X é proporcional à intensidade da cir-

culação convectiva, Y quantifica a diferença de temperatura entre as correntes ascendentes e descendentes, e Z representa o desvio do perfil vertical de temperatura em relação a um gradiente linear de condução (LORENZ, 1963). O comportamento dinâmico desse sistema é governado por três parâmetros adimensionais fundamentais, σ , r e β , que controlam a transição entre regimes estáveis e caóticos.

O parâmetro σ representa o número de Prandtl, definido pela razão entre a difusividade momento e a difusividade térmica do fluido. Fisicamente, ele determina a velocidade relativa com que o momento e o calor se difundem através do fluido; no modelo clássico, utiliza-se $\sigma = 10$ (LORENZ, 1963; GHIL; CHILDRESS, 1987).

O parâmetro r corresponde ao número de Rayleigh reduzido ($r = R/R_c$), no qual R_c representa o número de Rayleigh crítico necessário para o início da convecção térmica. Este termo atua como o principal parâmetro de controle do sistema. O número de Rayleigh (R) quantifica a instabilidade de uma camada de fluido devido à diferença de temperatura entre a base aquecida e o topo resfriado. Quando r é pequeno, o calor é transferido apenas por condução; contudo, à medida que r aumenta e ultrapassa a unidade, a flutuabilidade supera a resistência viscosa, iniciando o movimento convectivo. Para valores elevados, como $r = 28$ ou $r = 32$, o sistema entra em regime caótico, no qual a dissipação de energia já não é suficiente para estabilizar o escoamento em um estado estacionário (GHIL; CHILDRESS, 1987).

Por fim, β é um parâmetro geométrico associado à razão de aspecto da célula de convecção, comumente fixado em $8/3$ para representar a estrutura física simplificada proposta originalmente por Lorenz (1963).

2.2.2 Estabilidade e Análise dos Pontos Críticos

A análise da estabilidade dos pontos críticos serve para mapear os estados de equilíbrio do sistema e entender sob quais condições e limiares de parâmetros a dinâmica deixa de ser estável e transiciona para o comportamento oscilatório ou caótico. Esta abordagem constitui etapa fundamental para compreender o comportamento global das soluções. Esses pontos são obtidos ao igualar as Equações (1), (2) e (3) a zero. Identifica-se, inicialmente, o ponto crítico na origem, $P_0 = (0, 0, 0)$. Por meio da linearização do sistema e da análise dos autovalores da matriz Jacobiana associada, demonstra-se que P_0 é um atrator estável para $r < 1$ (GHIL; CHILDRESS, 1987).

Quando o número de Rayleigh reduzido excede a unidade ($r > 1$), ocorre uma bifurcação na qual P_0 se torna um ponto de sela instável. Nesse estágio, surgem dois novos pontos críticos simétricos, denominados P_1 e P_2 :

$$P_{1,2} = \left(\pm \sqrt{\beta(r-1)}, \pm \sqrt{\beta(r-1)}, r-1 \right) \quad (4)$$

Esses pontos representam estados de convecção estacionária. Para os valores

clássicos de $\sigma = 10$ e $\beta = 8/3$, os pontos P_1 e P_2 permanecem estáveis apenas enquanto r for menor que o valor crítico $r_c \approx 24,74$. Para $r > r_c$, esses pontos perdem a estabilidade por meio de uma bifurcação de Hopf. Esse fenômeno caracteriza a transição na qual um ponto de equilíbrio estável se torna instável e o sistema passa a exibir comportamento oscilatório persistente. No contexto do presente modelo, esse limiar marca a passagem de um regime dominado pela convergência para pontos fixos para um regime de comportamento muito mais complexo, no qual as trajetórias deixam de se estabilizar em estados estacionários (GHIL; CHILDRESS, 1987).

2.2.3 Divergência de Trajetórias e Regimes Dinâmicos

A investigação da previsibilidade no sistema é apresentada por Huang e Yang (1996) por meio da análise comparativa de trajetórias em dois regimes dinâmicos distintos. As autoras utilizam simulações numéricas para detalhar como a topologia do espaço de fases condiciona a evolução temporal do erro. O protocolo experimental fundamenta-se na integração das equações de estado a partir de uma condição inicial de referência ($X = 1,0$; $Y = 3,0$; $Z = 5,0$), comparando a evolução dessa trajetória de controle com uma trajetória submetida a uma perturbação inicial.

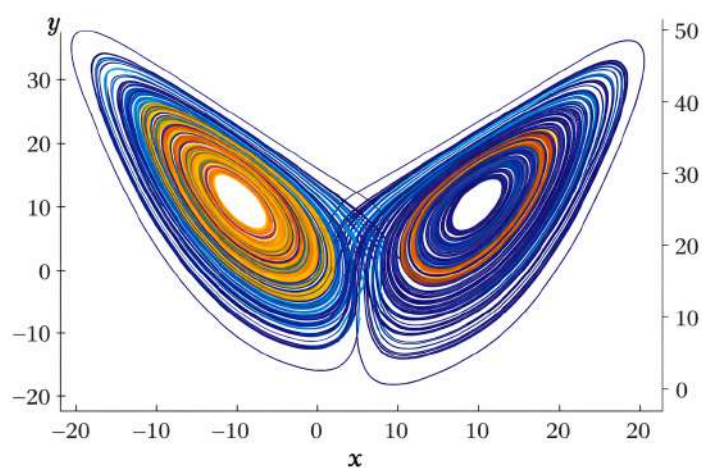
No regime não linear estável ($r = 10$), as soluções do sistema encontram-se sob a forte influência atratora dos pontos críticos estáveis P_1 e P_2 , definidos na Equação (4). Conforme descrito pelas autoras, ao introduzir uma perturbação nas condições iniciais de magnitude 10^{-1} , a divergência entre as trajetórias cresce de forma lenta e previsível (Figura 2b). Nesse estado, a dinâmica permanece confinada a uma região do espaço de fases na qual a linearização do modelo ainda constitui uma aproximação válida por intervalos relativamente longos de integração. Esse comportamento indica que a estabilidade estrutural do sistema atenua a propagação das incertezas iniciais.

Por outro lado, no regime caótico ($r = 32$), o sistema evolui sobre um atrator estranho, descrito por Boyce, DiPrima e Meade (2024) como uma estrutura geométrica de natureza aperiódica e alta complexidade. Huang e Yang (1996) descrevem que, partindo de condições iniciais idênticas às do regime estável, a mesma perturbação inicial desencadeia uma separação exponencial e abrupta das trajetórias (Figura 2a). Esse fenômeno ocorre em razão da instabilidade intrínseca do atrator: ao orbitar um dos lóbulos, a trajetória atinge regiões de sela nas quais pequenas diferenças na posição atual podem determinar se o sistema permanecerá no lóbulo vigente ou se será repellido em direção ao lóbulo oposto.

Essa alternância aperiódica entre os lóbulos caracteriza a perda de previsibilidade determinística. Huang e Yang (1996) ressaltam que, no regime caótico, a distância entre trajetórias vizinhas no espaço de fases é governada pelo maior expoente de Lyapunov (λ_1). Isso implica que qualquer incerteza na análise inicial, independentemente de sua magnitude, tende a dominar a solução após um tempo finito, cuja ordem de

grandeza é frequentemente associada a $t \approx 1/\lambda_1$ (HUANG; YANG, 1996). A estrutura bimodal do atrator, conhecida como “asas da borboleta”, impõe que, em horizontes temporais superiores a esse limite, o sistema apresente estados estatisticamente divergentes.

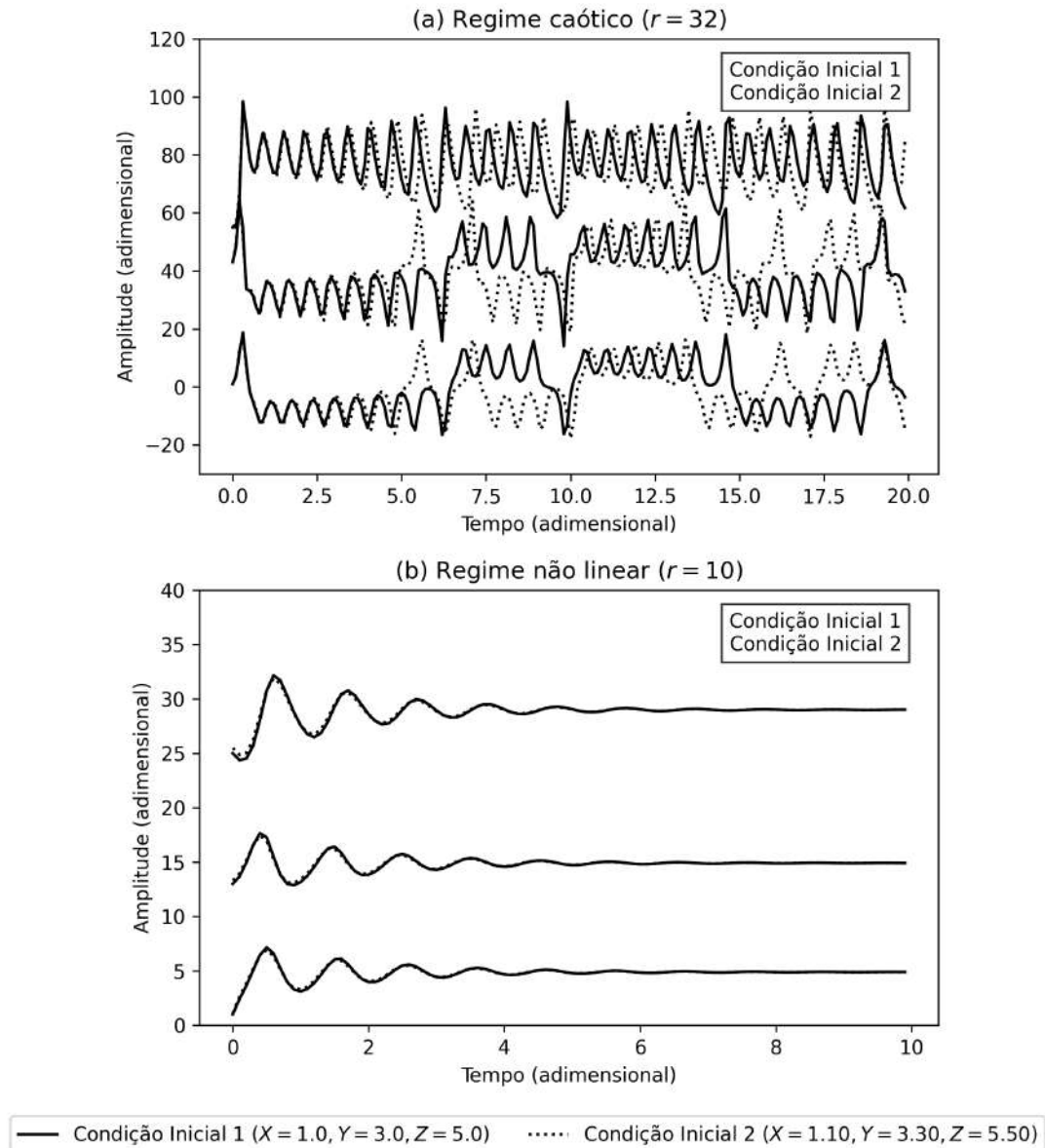
Figura 1 – Representação tridimensional do atrator estranho de Lorenz no espaço de fases (X, Y, Z) para o regime caótico ($r = 32$), evidenciando a estrutura bimodal conhecida como asas da borboleta.



Fonte: Autoria própria.

De acordo com Wilks (2011), essa bimodalidade cria um cenário crítico para a estatística da previsão. Como as trajetórias tendem a se concentrar preferencialmente nas regiões que circundam os pontos de equilíbrio instáveis P_1 e P_2 , a média aritmética das trajetórias, ou do *ensemble*, pode se posicionar na região central entre os dois lóbulos. Esse posicionamento resulta em um estado médio localizado em uma região de baixa densidade probabilística, pouco representativa da estrutura dinâmica mais provável do sistema. Essa característica reforça que a estabilidade e a confiabilidade das previsões dependem intrinsecamente do parâmetro de forçante térmica r .

Figura 2 – Divergência de trajetórias no modelo de Lorenz nos regimes (a) caótico ($r = 32$) e (b) não linear estável ($r = 10$). A simulação apresenta a evolução temporal das variáveis de estado (X, Y, Z) a partir da condição inicial de controle (linha sólida) e da condição perturbada em 10^{-1} (linha tracejada).



Fonte: Autoria própria a partir da metodologia descrita por Huang e Yang (1996).

A análise desse sistema é particularmente relevante para a presente pesquisa porque oferece um ambiente matemático controlado no qual a sensibilidade às condições iniciais, a multimodalidade das trajetórias e a limitação da média simples podem ser investigadas de forma explícita. Vale destacar que a aparente similaridade visual observada na evolução temporal das variáveis Y e Z no regime caótico (Figura 2a) não constitui mera coincidência, mas reflete o forte acoplamento dinâmico não linear do sistema; a variação temporal de Z depende diretamente do produto das amplitudes de X e Y , fazendo com que as órbitas alternadas em torno dos lóbulos do atrator manifestem comportamentos oscilatórios sincronizados e correspondentes no tempo.

Assim, o modelo fornece uma base conceitual apropriada para avaliar estratégias de representação de *ensembles* em regimes dinâmicos distintos.

2.3 Aprendizado de Máquina Não Supervisionado

O Aprendizado de Máquina (*Machine Learning*) consiste no desenvolvimento de métodos estatísticos e computacionais capazes de identificar relações, regularidades e estruturas a partir de amostras, permitindo realizar inferência ou previsão sem a necessidade de uma modelagem explícita baseada exclusivamente em princípios determinísticos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Dentro desse campo, o aprendizado não supervisionado caracteriza-se por lidar com conjuntos de dados nos quais não há variáveis de saída previamente definidas, tendo como objetivo principal descobrir propriedades intrínsecas à distribuição dos dados (BISHOP, 2006).

Diferentemente do aprendizado supervisionado, no qual o modelo é ajustado a partir de pares entrada-saída conhecidos, o aprendizado não supervisionado busca descrever a estrutura interna do conjunto de observações, podendo ser interpretado como um problema de modelagem de distribuições de probabilidade, redução de dimensionalidade ou identificação de agrupamentos naturais.

Uma das tarefas mais comuns dentro desse paradigma é o agrupamento de dados, ou *clustering*. O objetivo do agrupamento consiste em particionar um conjunto de observações em subconjuntos, de modo que elementos pertencentes ao mesmo grupo apresentem alta similaridade entre si, enquanto elementos de grupos distintos sejam tão diferentes quanto possível. Essa similaridade é normalmente definida por meio de métricas de distância no espaço de atributos, sendo a distância Euclidiana uma escolha frequente em problemas contínuos, embora outras métricas também possam ser utilizadas, dependendo da geometria do espaço de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Sob uma perspectiva probabilística, muitos métodos de aprendizado de máquina não supervisionado podem ser interpretados como técnicas de estimação de densidade, nas quais o objetivo consiste em identificar regiões do espaço de estados em que as observações se concentram com maior probabilidade. No contexto de algoritmos de agrupamento baseados em centróides, essa concentração de pontos pode ser representada por um número finito de centros ou representantes, permitindo uma representação sintética de conjuntos de dados de alta dimensionalidade (BISHOP, 2006; JAIN, 2010).

A capacidade de organizar informações com base exclusivamente na estrutura geométrica dos dados torna o aprendizado não supervisionado particularmente adequado à análise de sistemas dinâmicos complexos e multidimensionais. Nesses casos, a identificação de padrões recorrentes ou de regimes dinâmicos distintos pode ser

realizada sem a necessidade de rótulos prévios, o que favorece a aplicação de métodos de particionamento voltados à identificação de regimes e padrões recorrentes. No presente trabalho, esse interesse recai sobre a identificação de estruturas recorrentes nas trajetórias produzidas por *ensembles* em um sistema dinâmico caótico.

Entre os métodos de agrupamento mais utilizados nesse contexto, destaca-se o algoritmo *K-means*, amplamente empregado em problemas de particionamento de dados contínuos em razão de sua simplicidade computacional e de sua interpretabilidade. Em virtude de sua relevância para a presente pesquisa, esse método é detalhado na subseção a seguir.

2.3.1 Algoritmo K-means e Critérios de Agrupamento

O algoritmo *K-means* é um dos métodos mais utilizados para a realização de agrupamentos (*clustering*) em problemas de aprendizado de máquina não supervisionado. Seu objetivo consiste em particionar uma amostra de observações em K grupos distintos, de modo que os elementos pertencentes ao mesmo grupo apresentem alta similaridade entre si, enquanto elementos de grupos diferentes sejam tão distintos quanto possível. Esse processo é realizado por meio da minimização da variabilidade interna de cada grupo, representada por uma função de erro baseada na distância entre as observações e seus respectivos centróides (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; BISHOP, 2006).

Historicamente, o desenvolvimento do algoritmo decorreu de contribuições sucessivas ao longo da década de 1960. Um dos primeiros métodos relacionados foi apresentado por Forgy (1965), seguido pela formulação proposta por MacQueen (1967), que introduziu o termo *K-means* e descreveu uma abordagem baseada na atualização sequencial dos centróides à medida que novas observações eram processadas. Nesse método, cada nova observação era atribuída ao centróide mais próximo e, em seguida, o centróide correspondente era atualizado com base na média das observações pertencentes ao grupo.

Posteriormente, Lloyd (1982) apresentaram uma formulação alternativa baseada em processamento em lote (*batch*), que se tornou a versão mais difundida do algoritmo. Nessa abordagem, todas as observações são inicialmente atribuídas ao centróide mais próximo e, somente após essa etapa, os centróides são recalculados com base na média dos pontos pertencentes a cada grupo. Esse processo iterativo pode ser interpretado como um problema de quantização vetorial que busca minimizar a soma dos erros quadráticos dentro dos grupos.

Formalmente, o objetivo do algoritmo consiste em minimizar a função conhecida como Soma dos Quadrados Intra-Grupos (*Within-Cluster Sum of Squares* – WCSS):

$$WCSS = \sum_{j=1}^K \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (5)$$

onde C_j representa o conjunto de pontos pertencentes ao *cluster* j , x_i representa uma observação do conjunto de dados e μ_j corresponde ao centróide do *cluster* j , definido pela média aritmética das observações atribuídas a esse grupo.

Do ponto de vista geométrico, o *K-means* realiza uma partição do espaço de atributos conhecida como Tesselação de Voronoi. Cada centróide atua como o gerador de uma célula poligonal (ou hiperbólica em dimensões superiores), delimitando regiões em que qualquer ponto contido em seu interior está matematicamente mais próximo daquele centro do que de qualquer outro centro concorrente. Como o algoritmo realiza uma busca gulosa local através de descida de coordenadas, o formato das fronteiras de decisão é estritamente linear, o que implica que o método assume implicitamente que os agrupamentos naturais dos dados possuem uma simetria esférica e distribuição aproximadamente isotrópica no espaço amostral (BISHOP, 2006).

A minimização dessa função ocorre por meio de um procedimento iterativo composto por duas etapas principais: atribuição e atualização. Inicialmente, cada observação é atribuída ao centróide mais próximo, de acordo com uma métrica de distância, geralmente a distância Euclidiana. Em seguida, os centroides são recalculados como a média das observações pertencentes a cada grupo. Esse processo é repetido até que os centroides se estabilizem ou até que um critério de parada seja atingido (BISHOP, 2006; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

De forma geral, o funcionamento do algoritmo pode ser descrito por meio do seguinte ciclo operacional:

1. **Inicialização:** definição de K centroides iniciais no espaço de dados;
2. **Atribuição:** cada observação x_i é associada ao *cluster* cujo centróide μ_j minimiza a distância $\|x_i - \mu_j\|$;
3. **Atualização:** os centroides são recalculados como a média dos pontos pertencentes a cada *cluster*;
4. **Convergência:** as etapas anteriores são repetidas até que as atribuições dos pontos não se alterem ou até que a redução da função de erro se torne insignificante.

Apesar de sua simplicidade computacional e eficiência, o algoritmo apresenta limitações importantes. Uma delas está relacionada à necessidade de definir previamente o número de *clusters* K . A escolha inadequada desse parâmetro pode levar a agrupamentos pouco representativos da estrutura dos dados. Para auxiliar nessa decisão,

diferentes critérios de avaliação podem ser utilizados (JAIN, 2010; HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Um dos métodos mais conhecidos é o Método do Cotovelo (*Elbow Method*), cuja fundamentação conceitual de classificação baseada em quebras de variância remete ao trabalho pioneiro de Thorndike (1953), e que consiste em analisar a variação do WCSS em função do número de *clusters*. Nesse método, adaptado posteriormente para a validação do algoritmo geométrico de agrupamento, o valor ideal de K é frequentemente associado ao ponto de inflexão da curva, a partir do qual a redução da função de erro interna passa a ocorrer de forma menos acentuada.

Outra métrica amplamente utilizada para avaliar a coesão e a separação das partições geradas é o Coeficiente de Silhueta (*Silhouette Score*), proposto por Rousseeuw (1987). Esse índice mede simultaneamente a proximidade interna e a distância entre *clusters*, permitindo avaliar quão adequadamente cada observação foi alocada. Para cada ponto i , o coeficiente de silhueta é definido como

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

onde $a(i)$ representa a distância média entre o ponto i e todos os outros pontos pertencentes ao mesmo *cluster*, enquanto $b(i)$ representa a menor distância média entre o ponto i e os pontos pertencentes aos demais *clusters*. O valor de $s(i)$ varia no intervalo $[-1, 1]$, sendo que valores próximos de 1 indicam boa separação entre *clusters*, ao passo que valores próximos de 0 sugerem sobreposição entre grupos.

2.3.1.1 Otimização da Inicialização: o método *K-means++*

Uma limitação conhecida do algoritmo *K-means* está associada à escolha dos centroides iniciais, pois diferentes inicializações podem conduzir a soluções distintas em razão da presença de múltiplos mínimos locais na função objetivo. Para reduzir essa sensibilidade, Arthur e Vassilvitskii (2007) propuseram o algoritmo *K-means++*, que introduz um procedimento probabilístico para selecionar centroides iniciais mais bem distribuídos no espaço de dados.

Nesse método, o primeiro centroide é escolhido aleatoriamente entre as observações. A seleção dos centroides subsequentes ocorre de forma probabilística, onde a chance de qualquer ponto candidato ser escolhido é diretamente proporcional ao quadrado da distância até o centroide mais próximo que já tenha sido estabelecido. Formalmente, a probabilidade $P(x)$ de seleção de um ponto x pertencente ao conjunto total de observações X é dada por

$$P(x) = \frac{D(x)^2}{\sum_{x' \in X} D(x')^2} \quad (7)$$

onde $D(x)$ representa a distância entre o ponto em análise x e o centroide mais

próximo já selecionado, enquanto o termo x' no denominador atua como uma variável de indexação para realizar o somatório de todas as distâncias quadráticas de cada elemento contido no conjunto X . Esse procedimento favorece a escolha de centros iniciais mais espaçados, o que tende a melhorar a adequação do ajuste estrutural e a acelerar a convergência do algoritmo.

2.3.1.2 Exemplo de Aplicação em Sistemas Meteorológicos

Para ilustrar a viabilidade operacional deste método no processamento de dados geofísicos, cita-se o cenário clássico de previsão de trajetórias de ciclones extratropicais ou frentes frias a partir de modelos de grande escala de circulação geral. Quando um modelo operacional estendido gera um macro-ensemble de 50 ou 100 membros simulados para os próximos 10 dias, a dispersão física das coordenadas espaciais dos centros de baixa pressão tende a aumentar exponencialmente devido à não linearidade dos fluidos geofísicos.

Ao aplicar o algoritmo *K-means++* tomando como vetor de atributos as matrizes temporais de latitude e longitude dos centros de tempestade em cada instante de integração, o algoritmo é capaz de segregar as simulações em cenários mutuamente exclusivos. Se o Coeficiente de Silhueta apontar de forma estável para $K = 2$, o meteorologista obtém um diagnóstico binário: o *Cluster 1* pode concentrar membros que indicam um deslocamento rápido do sistema meteorológico em direção ao oceano, enquanto o *Cluster 2* pode agrupar trajetórias lentas continentais, associadas a bloqueios atmosféricos regionais. Em vez de calcular uma média aritmética simples entre essas duas tendências opostas, o que resultaria em uma trajetória média artificial localizada sobre o litoral, onde nenhum membro de fato simulou, a análise baseada em centroides preserva as duas estruturas dinâmicas alternativas perfeitamente plausíveis (ALHAMED; LAKSHMIVARAHAN; STENSRUD, 2002).

2.3.1.3 Dinâmica de Ensembles e Representação por Centroides

No contexto da previsão por conjuntos aplicada a sistemas caóticos, diferentes simulações são realizadas a partir de condições iniciais levemente perturbadas com o objetivo de amostrar o crescimento das incertezas e mapear a dispersão probabilística das trajetórias no espaço de fases. Tradicionalmente, uma forma usual de síntese da previsão consiste em utilizar a média simples das trajetórias do *ensemble*. Essa abordagem baseia-se na premissa de que erros individuais tendem a se compensar quando considerados de forma agregada, embora isso não assegure, em todos os casos, a representação mais aderente à trajetória de referência.

Entretanto, estudos na área de previsão numérica do tempo mostram que a média do *ensemble* pode produzir estados suavizados e, em algumas situações, fisicamente pouco plausíveis, especialmente quando os membros divergem para diferentes re-

giões do espaço de fases (TOTH; KALNAY, 1993; LEUTBECHER; PALMER, 2008). Nesses casos, a média aritmética pode resultar em um estado localizado em regiões de baixa densidade de probabilidade (WILKS, 2011).

Uma alternativa consiste em empregar técnicas de agrupamento para identificar subconjuntos de membros do *ensemble* que compartilham evoluções dinâmicas semelhantes. Nesse contexto, os centroides dos *clusters* podem ser utilizados como representantes de trajetórias predominantes do sistema, preservando características dinâmicas que podem ser atenuadas pela média simples. Ainda assim, essa representação deve ser compreendida como uma forma de síntese estrutural do conjunto, e não como garantia *a priori* de redução do RMSE em relação à média aritmética (ALHAMED; LAKSHMIVARAHAN; STENSRUD, 2002; JAIN, 2010).

Para avaliar quantitativamente o desempenho dessas previsões em relação à trajetória de referência, utiliza-se com frequência a métrica da Raiz do Erro Quadrático Médio (*Root Mean Square Error* – RMSE), definida como

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (x_{prev,i} - x_{ref,i})^2} \quad (8)$$

onde $x_{prev,i}$ representa o valor previsto pelo modelo, $x_{ref,i}$ corresponde ao valor observado ou de referência e M representa o número total de observações consideradas. O RMSE penaliza erros de maior magnitude e constitui uma das métricas mais utilizadas na avaliação de modelos de previsão em meteorologia e ciências atmosféricas (WILKS, 2011).

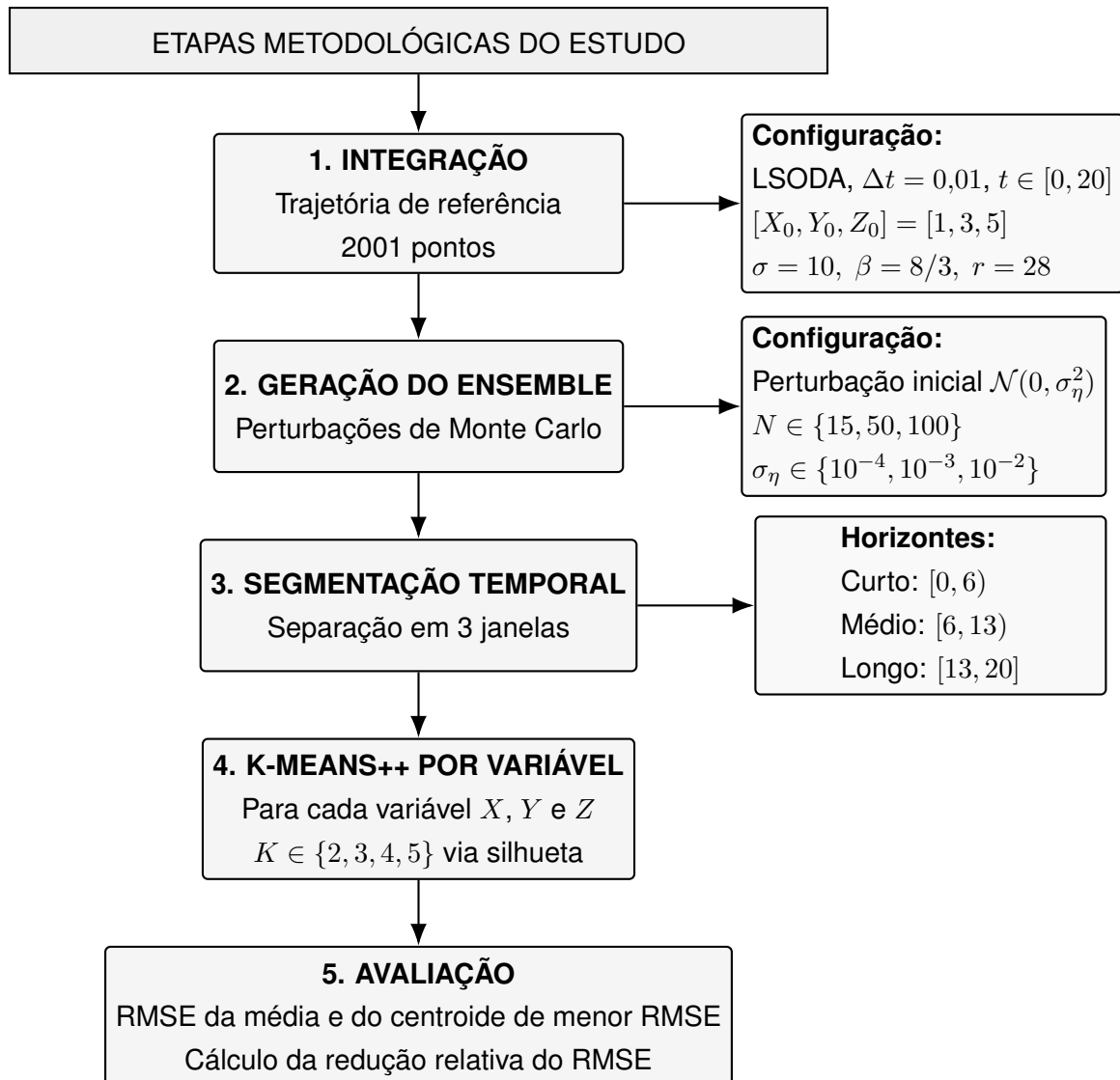
Dessa forma, o aprendizado não supervisionado fornece os fundamentos matemáticos necessários para investigar a organização geométrica das trajetórias de um *ensemble*. No contexto desta dissertação, o algoritmo *K-means++* é aplicado para identificar a assinatura topológica e a distribuição estrutural dos membros gerados pelas integrações numéricas, permitindo comparar a representação fornecida por centroides com aquela obtida pela média simples do conjunto, com base no comportamento das trajetórias e na avaliação do RMSE em relação à solução de referência (BISHOP, 2006; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAIN, 2010).

3 METODOLOGIA

Este capítulo descreve os procedimentos numéricos e estatísticos adotados para analisar, no sistema de Lorenz, diferentes formas de representação de previsões por conjuntos e comparar, em distintos cenários experimentais, a média aritmética do *ensemble* com trajetórias representativas obtidas por agrupamento via *K-means++*. A metodologia foi organizada em etapas sucessivas, que incluem a integração numérica da trajetória de referência, a geração dos membros perturbados, a segmentação temporal das trajetórias, o agrupamento segregado por variável em cada horizonte e a avaliação quantitativa dos estimadores por meio de métricas de erro.

A Figura 3 apresenta, de forma sintética, o fluxograma das etapas metodológicas empregadas neste estudo.

Figura 3 – Fluxograma das etapas metodológicas adotadas neste estudo.



Fonte: Autoria própria.

3.1 Integração Numérica do Modelo de Lorenz

A dinâmica temporal do sistema de Lorenz foi obtida por integração numérica do sistema de equações diferenciais ordinárias definido por Lorenz (1963). Para esse propósito, utilizou-se a função `odeint` da biblioteca `SciPy`, a qual emprega internamente o integrador LSODA, amplamente utilizado na resolução de sistemas não lineares e capaz de alternar automaticamente entre métodos adequados a regimes rígidos e não rígidos (HINDMARSH, 1983; PETZOLD, 1983).

As integrações foram realizadas no intervalo temporal adimensional $t \in [0, 20]$, com armazenamento das soluções em 2001 instantes igualmente espaçados, correspondendo a uma resolução temporal de saída de $\Delta t = 0,01$. Ressalta-se, entretanto, que esse valor representa o espaçamento entre os pontos armazenados da solução, e não

necessariamente o passo interno efetivo do integrador, uma vez que o LSODA utiliza controle adaptativo de passo durante o processo de integração.

A condição inicial de referência adotada foi

$$\mathbf{x}_0 = (X_0, Y_0, Z_0) = (1, 3, 5), \quad (9)$$

enquanto os parâmetros do sistema foram fixados nos valores clássicos

$$\sigma = 10, \quad \beta = \frac{8}{3}, \quad r = 28, \quad (10)$$

os quais caracterizam um regime caótico amplamente utilizado na literatura (SPARROW, 1982).

A trajetória obtida a partir dessa configuração foi considerada como trajetória de referência, isto é, a solução produzida a partir da condição inicial não perturbada. Essa trajetória foi utilizada como base para a avaliação do erro dos estimadores construídos a partir do *ensemble* (WILKS, 2011).

3.2 Geração do Ensemble

A representação da incerteza nas condições iniciais foi realizada por meio da construção de um *ensemble* de trajetórias perturbadas. Cada membro do conjunto foi gerado a partir da adição de uma perturbação aleatória gaussiana à condição inicial de referência. Neste trabalho, essa perturbação inicial será, doravante, referida também como ruído, sem mudança no significado adotado no estudo. Assim, a condição inicial do membro m foi definida por

$$\mathbf{x}_0^{(m)} = \mathbf{x}_0 + \boldsymbol{\eta}^{(m)}, \quad (11)$$

em que $\mathbf{x}_0^{(m)}$ representa a condição inicial do membro m do *ensemble* e $\boldsymbol{\eta}^{(m)}$ é um vetor aleatório com distribuição

$$\boldsymbol{\eta}^{(m)} \sim \mathcal{N}(\mathbf{0}, \sigma_\eta^2 I_3), \quad (12)$$

sendo I_3 a matriz identidade de ordem 3 e σ_η o desvio-padrão da perturbação inicial aplicada às componentes X , Y e Z . Essa formulação segue a lógica clássica da previsão por conjuntos, na qual perturbações nas condições iniciais são utilizadas para representar incertezas de análise e observação (LEITH, 1974; TOTH; KALNAY, 1993; KALNAY, 2003).

A escolha de uma perturbação gaussiana com média nula implica que, em média, não se introduz deslocamento sistemático da condição inicial de referência em nenhuma direção específica do espaço de estados. Em vez disso, a perturbação atua como mecanismo de dispersão inicial controlada em torno da trajetória de referência,

permitindo avaliar como pequenas diferenças impostas ao estado inicial evoluem sob a dinâmica do sistema.

Foram considerados três níveis de perturbação:

$$\sigma_\eta \in \{10^{-4}, 10^{-3}, 10^{-2}\}, \quad (13)$$

permitindo avaliar a sensibilidade do agrupamento e dos estimadores comparados a diferentes magnitudes de incerteza inicial. Além disso, o tamanho do *ensemble* foi variado segundo as configurações

$$N \in \{15, 50, 100\}, \quad (14)$$

de modo a investigar o efeito do tamanho do conjunto na identificação de padrões dinâmicos e na estabilidade dos agrupamentos (STENSRUD et al., 2000; WILKS, 2011).

Cada membro do conjunto foi integrado individualmente no mesmo intervalo temporal da trajetória de referência, preservando os mesmos parâmetros do sistema e a mesma malha temporal de saída. Dessa forma, as diferenças entre as trajetórias do *ensemble* decorrem, do ponto de vista do experimento, exclusivamente das perturbações introduzidas na condição inicial.

Com o objetivo de assegurar reprodutibilidade numérica e comparabilidade direta entre os cenários experimentais, foi adotada uma semente pseudoaleatória fixa (*seed* = 42). Para cada combinação entre tamanho do *ensemble* e magnitude de perturbação inicial, o *ensemble* foi gerado uma única vez a partir dessa mesma semente, sendo posteriormente analisado nos três horizontes temporais definidos no estudo. Dessa forma, curto, médio e longo prazo correspondem a recortes temporais distintos de uma mesma realização do *ensemble* in cada cenário experimental. Essa escolha permite que a comparação entre horizontes, magnitudes de perturbação e tamanhos de conjunto seja conduzida sobre uma base numérica estável e explicitamente reprodutível. Ressalta-se, contudo, que os resultados correspondem a uma realização específica do processo de perturbação aleatória, o que limita a generalização estatística dos resultados e deve ser considerado na interpretação do alcance metodológico do estudo.

3.3 Definição dos Horizontes Temporais

A análise foi segmentada em três horizontes temporais, definidos operacionalmente a partir dos intervalos de tempo utilizados no processamento numérico:

- **Curto Prazo:** $0 \leq t < 6$;
- **Médio Prazo:** $6 \leq t < 13$;

- **Longo Prazo:** $13 \leq t < 20$.

Esses intervalos foram escolhidos com o objetivo de distinguir fases com diferentes padrões de dispersão e organização das trajetórias: uma fase inicial de menor separação entre os membros, um estágio intermediário em que a divergência entre trajetórias se torna mais evidente e um estágio tardio em que a evolução do conjunto passa a refletir de forma mais intensa a geometria global do atrator (SPARROW, 1982; WILKS, 2011). Trata-se, portanto, de janelas operacionais definidas para contrastar regimes progressivos de dispersão, adotadas para fins de comparação metodológica.

No processamento computacional, esses horizontes foram implementados a partir de recortes sobre a malha temporal total. Considerando a discretização adotada, os intervalos corresponderam aos subconjuntos de índices $[0, 600)$, $[600, 1300)$ e $[1300, 2000)$, respectivamente. Dessa forma, a análise quantitativa e o agrupamento foram realizados separadamente em cada uma dessas janelas, permitindo comparar o desempenho dos estimadores em diferentes estágios da evolução dinâmica do sistema.

3.4 Estrutura dos Dados para Agrupamento

O agrupamento foi realizado de maneira independente e segregada para cada uma das variáveis de estado do sistema de Lorenz, isto é, X , Y e Z . Para cada cenário experimental definido por tamanho do *ensemble*, nível de perturbação inicial, horizonte temporal e variável do sistema, cada membro foi representado por um vetor temporal contendo os valores simulados naquele intervalo. Assim, para um dado horizonte h , o conjunto de dados submetido ao agrupamento assumiu a forma matricial

$$\mathbf{E}^{(h)} \in \mathbb{R}^{N \times T_h}, \quad (15)$$

em que N representa o número de membros do *ensemble* e T_h o número de instantes pertencentes ao horizonte temporal analisado.

Desse modo, o agrupamento não foi realizado sobre estados instantâneos isolados, mas sobre a forma temporal das trajetórias em cada janela de tempo. Essa escolha preserva a evolução temporal de cada simulação dentro do horizonte considerado, permitindo que a comparação entre os membros seja feita com base no comportamento da trajetória ao longo da janela, e não apenas em um único instante. Essa escolha é consistente com o uso do agrupamento como ferramenta para identificar trajetórias com comportamento dinâmico semelhante em conjuntos de simulações (ALHAMED; LAKSHMIVARAHAN; STENSRUD, 2002; JAIN, 2010).

Ao mesmo tempo, a opção pelo particionamento segregado implica que a correlação multivariada simultânea entre as componentes X , Y e Z não é explorada

de forma combinada no algoritmo principal. Essa estratégia metodológica justifica-se pelo objetivo central de investigar, sob a perspectiva individual de cada coordenada, o comportamento dos centroides em relação à média aritmética simples do conjunto. Adicionalmente, o tratamento isolado de cada componente elimina a necessidade de transformações prévias de padronização, impedindo que disparidades nas ordens de magnitude e nas escalas de variabilidade das variáveis de estado mascarem a identificação de padrões geométricos relevantes durante a convergência dos grupos.

3.5 Seleção do Número de Grupos

A determinação do número de grupos K foi realizada de forma automática por meio do Coeficiente de Silhueta (*Silhouette Score*), calculado para valores de

$$K \in \{2, 3, 4, 5\}. \quad (16)$$

Para cada valor testado, ajustou-se um modelo *K-means++* e calculou-se o índice médio de silhueta. O número de grupos selecionado foi aquele que maximizou esse índice:

$$K^* = \arg \max_{K \in \{2,3,4,5\}} S(K), \quad (17)$$

em que $S(K)$ denota o valor médio do coeficiente de silhueta para uma partição com K grupos Rousseeuw (1987).

O Coeficiente de Silhueta foi adotado por fornecer uma medida simultânea da coesão interna e da separação entre grupos. Para cada associação ou observação i , calcula-se inicialmente $a(i)$, que corresponde à distância média entre essa observação e as demais pertencentes ao mesmo grupo. Em seguida, calcula-se $b(i)$, definido como a menor distância média entre a observação i e os pontos pertencentes aos outros grupos. Com base nessas duas quantidades, obtém-se

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \quad (18)$$

O valor de $s(i)$ pertence ao intervalo $[-1, 1]$. Valores próximos de 1 indicam que a observação está bem alocada em seu grupo e distante dos grupos vizinhos; valores próximos de 0 indicam que a observação se encontra em região intermediária entre grupos; e valores negativos sugerem possível alocação inadequada. O valor médio de silhueta para uma dada partição resume, portanto, o equilíbrio entre compactação intragrupo e separação intergrupos.

A restrição inferior $K = 2$ foi adotada por razões metodológicas e pela própria natureza do Coeficiente de Silhueta, o qual não é definido para soluções com $K = 1$, uma vez que sua formulação depende da comparação entre distâncias intra e inter-

grupos. Já o limite superior $K = 5$ foi escolhido para evitar fragmentação excessiva em conjuntos de baixa cardinalidade e, ao mesmo tempo, permitir a detecção de possíveis subestruturas em cenários de maior heterogeneidade. Como o menor *ensemble* considerado neste estudo possui $N = 15$ membros, a ampliação do intervalo de busca para valores muito superiores de K poderia conduzir a partições excessivamente fragmentadas, com grupos muito pequenos e menor interpretabilidade dinâmica. Dessa forma, o intervalo $K \in \{2, 3, 4, 5\}$ foi adotado como compromisso entre flexibilidade para detectar múltiplos padrões de trajetória e preservação de grupos com significado analítico.

3.6 Aplicação do Algoritmo K-means++

Uma vez definido o número ótimo de grupos K^* , aplicou-se o algoritmo *K-means++* à matriz de dados correspondente à variável e ao horizonte temporal em análise. O *K-means++* constitui uma variante do *K-means* clássico que busca melhorar a etapa de inicialização dos centroides, reduzindo a probabilidade de escolha de centros iniciais excessivamente próximos entre si e favorecendo centroides iniciais mais bem distribuídos no espaço de atributos (ARTHUR; VASSILVITSKII, 2007).

O procedimento inicia-se com a escolha aleatória de uma observação do conjunto de dados para compor o primeiro centroide. Em seguida, para cada observação subsequente, calcula-se a distância até o centroide já selecionado mais próximo. Essas distâncias são utilizadas para construir uma distribuição de probabilidade em que observações mais afastadas dos centros já definidos possuem maior chance de serem escolhidas como novos centroides iniciais. Formalmente, a probabilidade de seleção de um ponto x pertencente ao conjunto total de observações $\mathcal{X}$ é dada por

$$P(x) = \frac{D(x)^2}{\sum_{x' \in \mathcal{X}} D(x')^2}, \quad (19)$$

em que $D(x)$ representa a distância entre o ponto em análise x e o centroide já escolhido mais próximo, enquanto o termo x' no denominador atua como uma variável de indexação para realizar o somatório de todas as distâncias quadráticas de cada elemento contido no conjunto $\mathcal{X}$. O uso do quadrado da distância faz com que regiões ainda pouco representadas do espaço de dados tenham maior probabilidade de contribuir para a inicialização dos grupos.

Após a seleção dos K^* centroides iniciais, o algoritmo passa a seguir o ciclo iterativo característico do *K-means*. Na primeira etapa, cada observação é atribuída ao centroide mais próximo segundo a distância Euclidiana. Em seguida, cada centroide é recalculado como a média aritmética das observações pertencentes ao respectivo grupo. A partir dessa atualização, uma nova etapa de atribuição é realizada, associ-

ando novamente cada observação ao centroide mais próximo. Esse processo iterativo de atribuição e atualização prossegue até que as alocações se estabilizem ou até que seja satisfeito o critério interno de convergência do algoritmo.

No presente estudo, cada observação corresponde a uma trajetória temporal estruturada por meio de uma análise individualizada por componente de estado de um membro do *ensemble* em uma janela específica. Assim, o processo de particionamento não é realizado sobre estados instantâneos isolados, mas sobre trajetórias contínuas completas, mapeadas independentemente para cada coordenada do sistema e representadas como vetores no espaço de atributos. Nessa formulação, os centroides obtidos pelo *K-means++* podem ser interpretados como trajetórias representativas associadas aos padrões mais recorrentes identificados em cada cenário experimental.

Em cada ajuste, utilizaram-se 10 inicializações independentes ($n_{init}=10$), preservando-se a semente fixa para assegurar reprodutibilidade e comparabilidade entre os experimentos. O emprego de múltiplas inicializações visa reduzir a sensibilidade do método à escolha inicial dos centroides e favorecer maior estabilidade da partição final, uma vez que o *K-means* poderá convergir para diferentes mínimos locais em função das condições iniciais do processo iterativo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.7 Definição do Centroide de Menor RMSE

Após a partição do conjunto em K^* grupos, obteve-se um conjunto de centroides, cada qual representando uma trajetória dentro do horizonte temporal analisado. Para fins de avaliação metodológica, cada centroide foi comparado à trajetória de referência correspondente ao mesmo intervalo de tempo por meio da métrica RMSE.

O centroide selecionado para comparação foi então definido como aquele que apresentou o menor erro quadrático médio em relação à referência:

$$c^* = \arg \min_{c_k} RMSE(c_k, \mathbf{x}_{ref}), \quad (20)$$

em que c_k representa o k -ésimo centroide e $\mathbf{x}_{ref}$ a trajetória de referência no horizonte considerado.

Essa estratégia permite verificar se a estrutura de agrupamento produzida pelo método contém trajetórias representativas mais aderentes à referência do que a média simples do conjunto, permitindo analisar o potencial representativo dos centroides no contexto controlado do experimento (ALHAMED; LAKSHMIVARAHAN; STENSRUD, 2002; WILKS, 2011). Como a seleção é feita a posteriori com base no RMSE, ela deve ser interpretada como parte do procedimento de avaliação adotado neste estudo.

3.8 Estimadores Comparados

Dois estimadores principais foram comparados ao longo deste estudo:

1. **Média aritmética do *ensemble***, obtida pela média ponto a ponto entre os membros do conjunto;
2. **Centroide de menor RMSE**, definido como o centroide que apresentou o menor erro em relação à trajetória de referência no horizonte correspondente.

A média aritmética do *ensemble* foi calculada por

$$\bar{x}(t) = \frac{1}{N} \sum_{m=1}^N x^{(m)}(t), \quad (21)$$

em que $x^{(m)}(t)$ representa o valor da variável de interesse para o membro m no instante t . Esse estimador corresponde à abordagem clássica de síntese de previsões por conjuntos (LEUTBECHER; PALMER, 2008; WILKS, 2011).

A comparação entre esses dois estimadores foi adotada por permitir o contraste direto entre uma forma tradicional de representação do *ensemble*, amplamente utilizada na literatura, e uma representação baseada em trajetórias representativas extraídas da estrutura interna do conjunto por meio de agrupamento.

3.9 Métricas de Avaliação

O desempenho dos estimadores foi avaliado por meio da Raiz do Erro Quadrático Médio (*Root Mean Square Error* – RMSE), definida por

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (x_{est,i} - x_{ref,i})^2}, \quad (22)$$

em que $x_{est,i}$ representa o valor estimado pelo método avaliado, $x_{ref,i}$ corresponde ao valor da trajetória de referência e M é o número de pontos temporais considerados no horizonte em análise (WILKS, 2011).

O RMSE foi adotado por fornecer uma medida direta do afastamento entre a trajetória estimada e a trajetória de referência, penalizando com maior intensidade discrepâncias de maior magnitude. Como o objetivo desta pesquisa consiste em comparar a aderência relativa entre diferentes representações do *ensemble* e a solução de referência, essa métrica oferece uma base objetiva e de interpretação direta para o contraste entre os estimadores.

Com base nessa métrica, calculou-se também a redução relativa do RMSE do

centroide de menor RMSE em relação à média aritmética do *ensemble*, segundo

$$\Delta\% = \left(\frac{RMSE_{\text{média}} - RMSE_{\text{centroide}}}{RMSE_{\text{média}}} \right) \times 100. \quad (23)$$

Valores positivos de $\Delta\%$ indicam desempenho relativo favorável ao centroide em relação à média do *ensemble*, enquanto valores negativos indicam desempenho relativo desfavorável. Essa medida foi utilizada para quantificar, em termos percentuais, a diferença entre os dois estimadores em cada cenário experimental.

Cabe observar que, em cenários operacionais nos quais o valor de $RMSE_{\text{média}}$ se aproxima de zero, padrão característico de horizontes de curtíssimo prazo, o indicador $\Delta\%$ torna-se matematicamente suscetível a instabilidades numéricas decorrentes do efeito de escala no denominador. Nessas condições, oscilações infinitesimais na diferença absoluta entre os estimadores são artificialmente amplificadas em termos percentuais, demandando que a análise da variação relativa seja estritamente vinculada aos valores absolutos correspondentes para evitar distorções na interpretação da real magnitude do ganho preditivo.

3.10 Análise Geométrica Complementar da Dispersão

Além do agrupamento segregado por componente e horizonte temporal, foi realizada uma análise geométrica complementar da configuração espacial do *ensemble* no instante final de cada horizonte. Essa etapa teve como objetivo oferecer uma visualização adicional da dispersão dos membros, permitindo representar, em projeções bidimensionais nos planos XY , XZ e YZ , a configuração final do *ensemble* em cada cenário experimental.

Para cada horizonte temporal, os estados tridimensionais finais dos membros foram considerados no espaço (X, Y, Z) , sendo também representados, nesse mesmo instante, a média aritmética do *ensemble*, a trajetória de referência e os centroides associados à análise principal para a variável em consideração. Dessa forma, as projeções bidimensionais foram utilizadas como recurso complementar de visualização da dispersão final do conjunto, preservando a correspondência com os cenários analisados nos gráficos temporais.

Nessa etapa, a média aritmética do *ensemble*, a trajetória de referência e os centroides finais foram representados graficamente, mantendo-se a mesma convenção de identificação utilizada ao longo da análise. Essa análise teve caráter complementar e ilustrativo, contribuindo para a visualização da dispersão espacial do *ensemble* ao final de cada horizonte (SPARROW, 1982; WILKS, 2011).

4 RESULTADOS E DISCUSSÕES

Neste capítulo, analisa-se o desempenho do algoritmo de agrupamento *K-means++*, aplicado por meio de uma análise individualizada por componente de estado às componentes do sistema, em contraste com a média aritmética simples do *ensemble*. A técnica é avaliada sob diferentes configurações de dimensão do *ensemble* ($N = 15, 50$ e 100), onde cada membro representa uma amostragem estocástica das trajetórias possíveis no espaço de fase.

As simulações foram segmentadas nos horizontes temporais de curto, médio e longo prazo, considerando três níveis de intensidade de perturbação nas condições iniciais, definidos pelos desvios padrão $\sigma = 10^{-4}, 10^{-3}$ e 10^{-2} . Os horizontes temporais adotados foram: curto prazo, no intervalo $0 \leq t < 6$; médio prazo, no intervalo $6 \leq t < 13$; e longo prazo, no intervalo $13 \leq t < 20$.

4.1 Análise do *Ensemble* referente à Média Aritmética Simples

A análise do comportamento dinâmico do *ensemble* revela as limitações intrínsecas da média aritmética simples à medida que a dispersão entre os membros evolui. Este diagnóstico fundamenta o referencial de desempenho necessário para validar os resultados obtidos via agrupamento por *K-means++*, evidenciando cenários em que a média simples perde a capacidade de acompanhar, com a mesma precisão, a distribuição do conjunto em relação à trajetória de referência.

4.1.1 Convergência e Sobreposição no Curto Prazo

Um padrão identificado nas simulações é a convergência das trajetórias no horizonte de curto prazo, no intervalo $0 \leq t < 6$. Nesta fase inicial, as perturbações nas condições iniciais ainda não sofreram a amplificação característica da natureza caótica do sistema. Como os membros apresentam evolução dinâmica análoga, a média simples e a trajetória de referência mantêm alta proximidade. Consequentemente, as métricas de tendência central mostram-se suficientes para representar o estado do sistema neste estágio.

No cenário com 100 membros e perturbação inicial de 10^{-2} , observa-se elevada sobreposição entre as trajetórias das variáveis X , Y e Z , evidenciando a estabilidade inicial do sistema de Lorenz. Neste estágio, a dispersão do *ensemble* é mínima, o que justifica a boa aderência da média aritmética simples antes do aumento da divergência entre os membros.

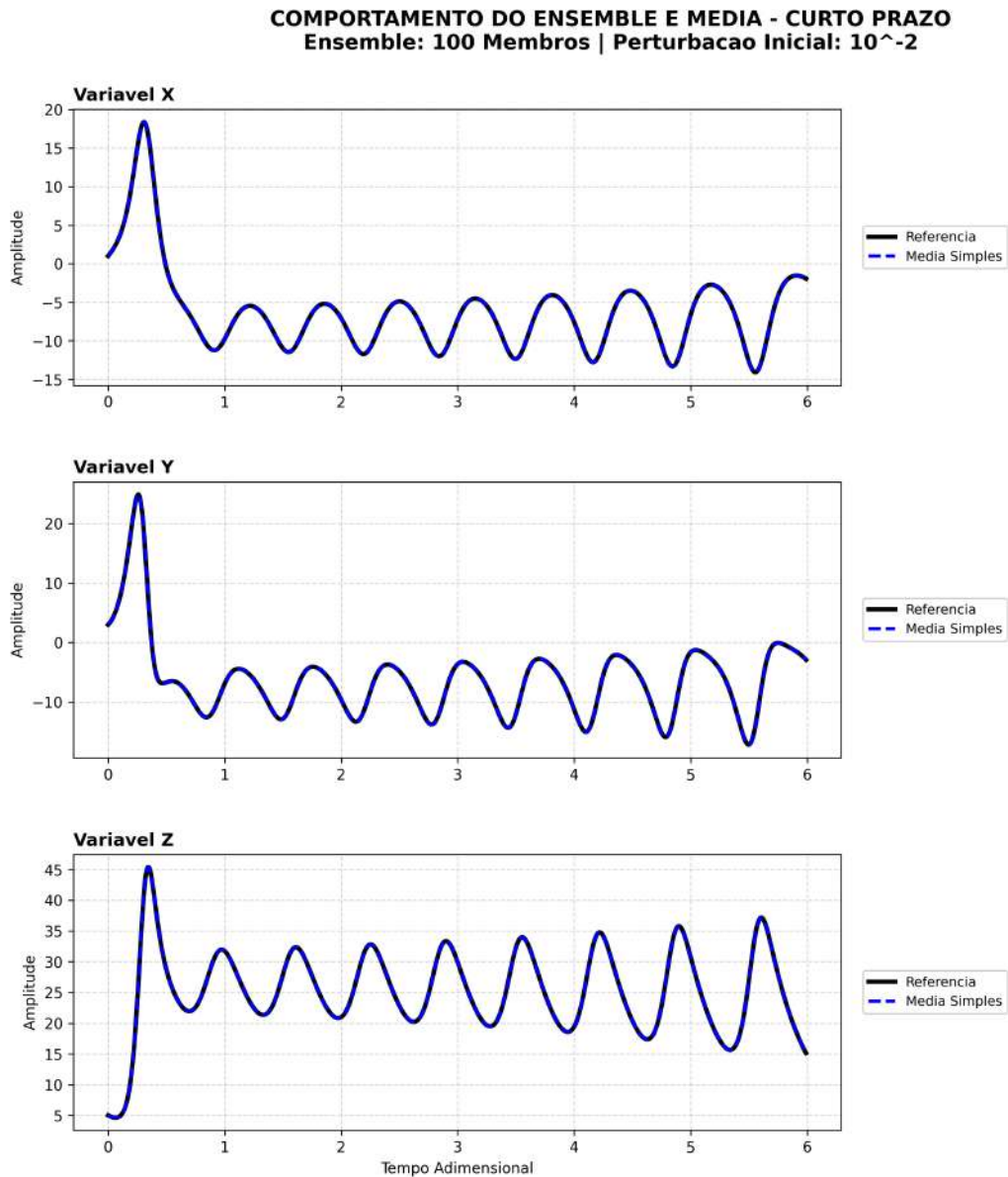


Figura 4 – Comportamento das variáveis X , Y e Z no curto prazo ($0 \leq t < 6$) para o cenário de 100 membros e ruído de 10^{-2} . As trajetórias dos membros apresentam alta convergência em relação à referência, evidenciando a baixa dispersão do *ensemble* nesse intervalo.

Fonte: Autoria própria.

4.1.2 Divergência e a limitação da Média no Médio Prazo

A análise do intervalo de médio prazo, entre $6 \leq t < 13$, mostra que o crescimento da dispersão compromete a capacidade da média aritmética simples de representar,

com a mesma precisão, a estrutura do *ensemble*. O desvio crescente em relação à trajetória de referência ocorre no momento em que os membros passam a ocupar regiões distintas do espaço de fase. Neste estágio, a transição para métodos de agrupamento passa a ser relevante, pois o conjunto já não se encontra tão concentrado quanto no curto prazo.

No cenário com 100 membros e perturbação inicial de 10^{-2} , observa-se que a média simples deixa de acompanhar com a mesma proximidade a distribuição das trajetórias. Sob o efeito da dinâmica caótica, o *ensemble* sofre um processo de cisão espacial em suas componentes X e Y , forçando os membros a se organizarem em regiões distintas e bimodais do espaço de fase (as asas do atrator). Ao computar o centro geométrico global desse sistema bifurcado, a média simples gera uma trajetória linear artificialmente suavizada que transita por zonas de baixa densidade probabilística, representando estados físicos irreais e evidenciando a limitação da média simples em resumir, sozinha, a estrutura do *ensemble* nesse intervalo.

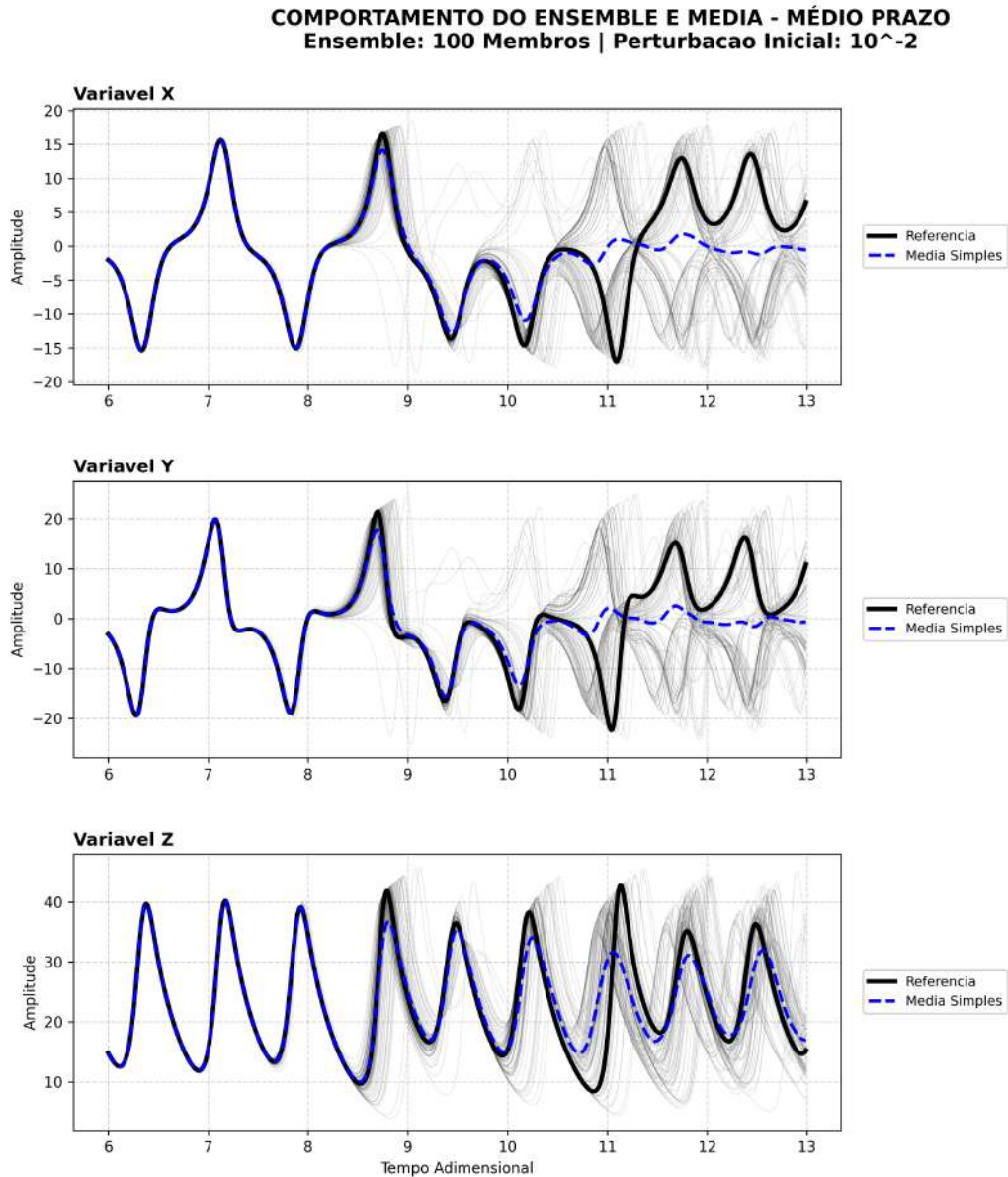


Figura 5 – Comportamento das variáveis X , Y e Z no médio prazo ($6 \leq t < 13$) para o cenário de 100 membros e ruído de 10^{-2} . O aumento da dispersão evidencia a limitação da média aritmética simples em representar a estrutura do *ensemble* nesse intervalo.

Fonte: Autoria própria.

4.1.3 Saturação da Dispersão no Longo Prazo

No horizonte de longo prazo, entre $13 \leq t < 20$, as simulações demonstram que o *ensemble* atinge um estado de saturação generalizada do erro. Independentemente da configuração da perturbação inicial, os membros distribuem-se de forma assíncrona, expandindo-se até preencher a escala de magnitude do próprio atrator de Lorenz. A informação associada à fase inicial dissipa-se completamente sob a ação do regime puramente caótico, de modo que os membros se distribuem de maneira aleatória sobre a geometria assintótica do sistema. Nesse estágio avançado, observa-se que o caos impera em todos os cenários experimentais avaliados. Diante disso, o de-

talhamento analítico minucioso dessa janela temporal torna-se secundário, uma vez que a perda total de previsibilidade determinística degrada de forma homogênea a capacidade de representação de estados instantâneos do modelo por qualquer operador preditivo.

No cenário com 100 membros e perturbação inicial de 10^{-4} , as trajetórias de X , Y e Z apresentam dispersão acentuada, evidenciando a saturação estatística do regime já estabelecido nesse intervalo.

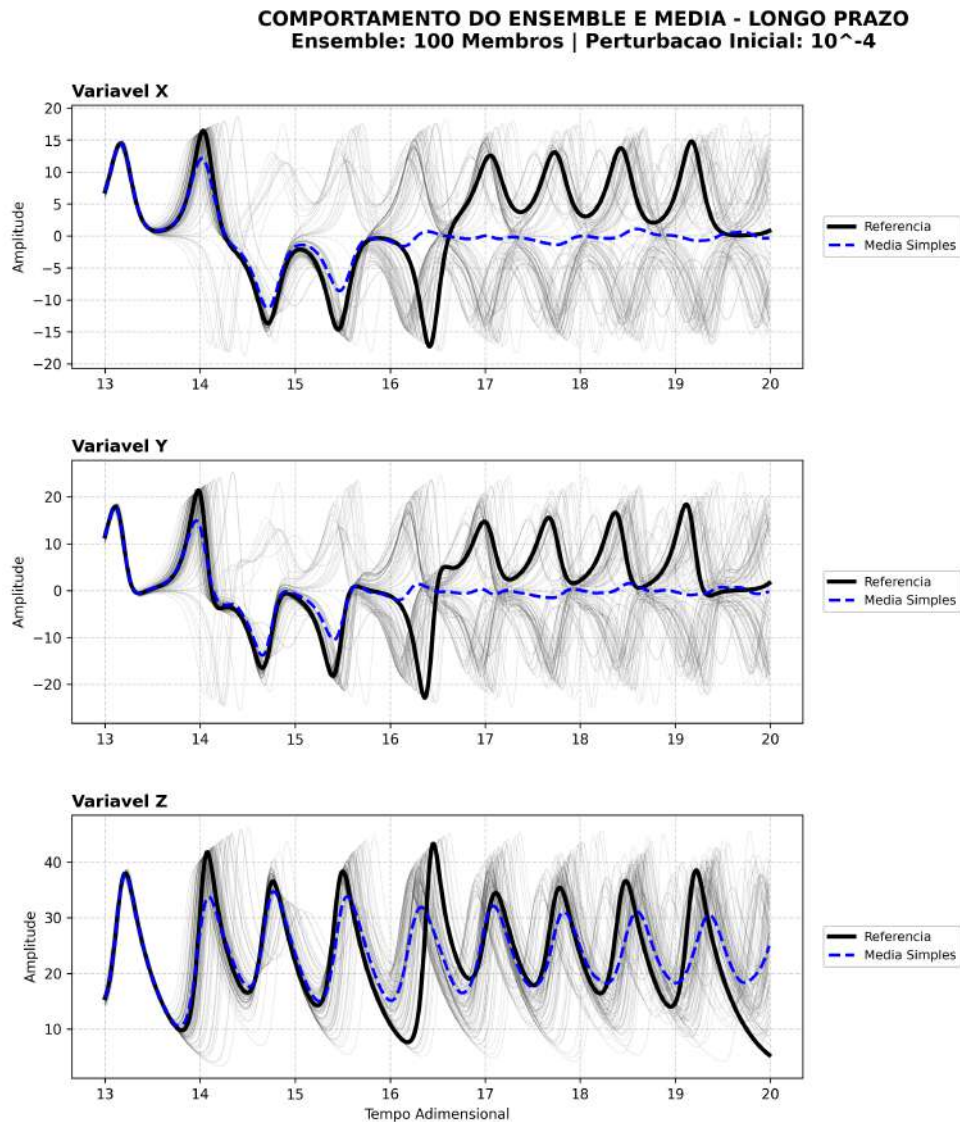


Figura 6 – Comportamento das variáveis X , Y e Z no longo prazo ($13 \leq t < 20$) para o cenário de 100 membros e ruído de 10^{-4} . A maior dispersão entre as trajetórias templates evidencia a limitação da média aritmética simples em representar o comportamento do *ensemble* nesse intervalo.

Fonte: Autoria própria.

4.2 Análise pelos Horizontes Temporais utilizando *K-means++*

Nesta seção, apresentam-se os resultados da integração temporal segmentados em três horizontes distintos: curto, médio e longo prazo. A comparação entre a média aritmética simples e o centróide de menor RMSE obtido pelo *K-means++* permite avaliar em quais situações o agrupamento identificou trajetórias mais próximas da referência para as variáveis X , Y e Z .

Antes da análise por subseção, os resultados gerais por horizonte temporal mostram que o centróide de menor RMSE superou a média aritmética simples em 18 dos 27 casos no curto prazo (66,67%), em 24 dos 27 casos no médio prazo (88,89%) e em 13 dos 27 casos no longo prazo (48,15%), totalizando 55 casos em que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples, em 81 experimentos (67,90%). Esses valores indicam que o desempenho do agrupamento depende intrinsecamente do horizonte temporal considerado e do nível de dispersão geométrica das trajetórias.

Nas tabelas a seguir, os percentuais positivos indicam que o centróide de menor RMSE apresentou valor menor que o da média aritmética simples em relação à trajetória de referência. Os percentuais negativos indicam que a média aritmética simples apresentou valor menor. Esses percentuais foram calculados a partir da diferença relativa entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE, tomando como referência o valor do RMSE da média aritmética simples.

Tabela 1 – Distribuição geral dos casos em que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples, por horizonte temporal.

Horizonte	Casos	Percentual
Curto prazo ($0 \leq t < 6$)	18 de 27	66,67%
Médio prazo ($6 \leq t < 13$)	24 de 27	88,89%
Longo prazo ($13 \leq t < 20$)	13 de 27	48,15%
Total	55 de 81	67,90%

Fonte: Autoria própria.

4.2.1 Curto Prazo ($0 \leq t < 6$)

No estágio inicial da integração, as trajetórias de todos os membros apresentam elevada convergência e reduzida dispersão em relação à trajetória de referência. As perturbações introduzidas nas condições iniciais ainda não foram amplificadas de forma suficiente para produzir uma separação mais nítida entre os membros, de modo que as trajetórias permanecem pouco dispersas nesse intervalo. Neste cenário, a média simples continua sendo uma estimativa estável do conjunto.

No entanto, os resultados atuais mostram que o comportamento do agrupamento

no curto prazo não foi o mesmo para todos os tamanhos de *ensemble*. Para os experimentos com 15 membros, os percentuais permaneceram positivos em todos os níveis de perturbação para as três variáveis, com valores próximos de 61%. Para 50 membros, os resultados também permaneceram positivos, com valores próximos de 76% a 78%. Já nos experimentos com 100 membros, os percentuais passaram a ser negativos em todos os níveis de perturbação e nas três variáveis, com valores bastante elevados em módulo, como $-309,03\%$ para X , $-307,88\%$ para Y e $-307,10\%$ para Z no cenário com perturbação inicial de 10^{-2} .

Esses valores devem ser interpretados estritamente em conjunto com a pequena magnitude dos RMSE absolutos nesse horizonte, uma vez que o indicador $\Delta\%$ torna-se altamente suscetível a flutuações numéricas severas (efeito de escala) quando o denominador se aproxima de zero. Pequenas diferenças infinitesimais em termos absolutos entre os estimadores acabam sofrendo uma amplificação artificial em termos percentuais, sem que isso represente uma perda real de qualidade física no ajuste estrutural.

Assim, nesta fase, não se pode afirmar que o agrupamento seja sempre desfavorável nem que seja sempre vantajoso. O que os resultados indicam é que, enquanto conjuntos com 15 e 50 membros apresentaram reduções positivas, o conjunto com 100 membros mostrou essa sensibilidade em relação à média simples. Nesse intervalo, a elevada convergência do *ensemble* ainda favorece uma representação mais estável pela média aritmética simples em várias configurações.

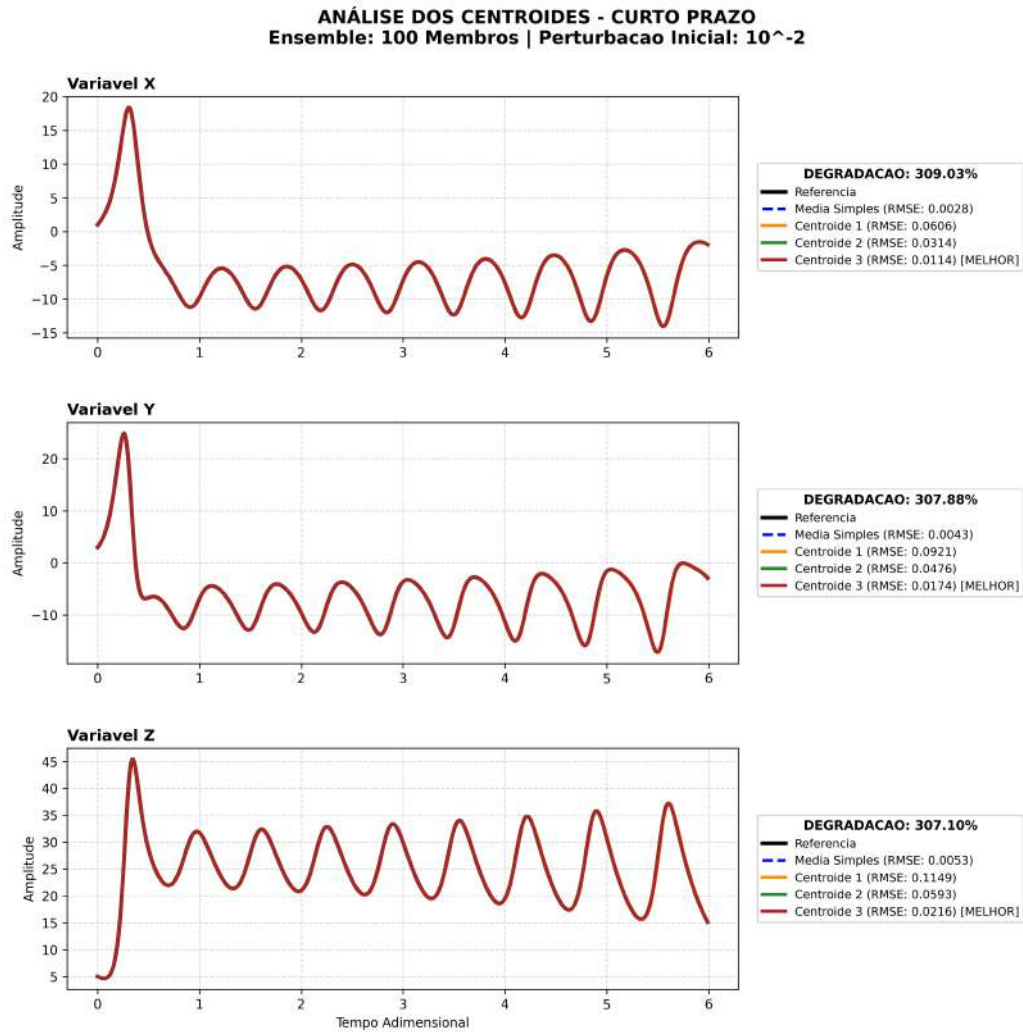


Figura 7 – Comportamento das variáveis X , Y e Z no curto prazo ($0 \leq t < 6$) para o cenário de 100 membros e perturbação de 10^{-2} , com aplicação do K -means++. As trajetórias apresentam alta convergência espacial, condição em que a média aritmética simples permanece estável e o agrupamento não apresenta vantagem nessa configuração específica.

Fonte: Autoria própria.

Tabela 2 – Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE no curto prazo ($0 \leq t < 6$), no cenário com ruído de 10^{-2} .

Número de membros	Variável X	Variável Y	Variável Z
15	61,91%	61,36%	61,43%
50	76,75%	78,34%	78,49%
100	-309,03%	-307,88%	-307,10%

Fonte: Autoria própria.

4.2.2 Médio Prazo ($6 \leq t < 13$)

O horizonte de médio prazo configurou-se como o intervalo em que o agrupamento apresentou melhor desempenho. Os resultados indicam que, neste intervalo, o en-

semble atinge uma janela de dispersão intermediária em que as trajetórias já se encontram suficientemente separadas para que o *K-means++* identifique a assinatura estrutural das ramificações, enquanto a dispersão ainda não alcançou o nível de saturação observado no longo prazo. Nesse contexto, a média simples perde capacidade de resumir adequadamente a distribuição do conjunto ao criar estados fisicamente inconsistentes com a dinâmica do modelo, e o agrupamento passa a fornecer centróides que preservam a integridade e a amplitude das trajetórias reais, gerando um expressivo ganho de previsibilidade em maior número de casos.

Esse comportamento aparece de forma clara em cenários com maior separação entre as trajetórias. No caso de 100 membros e perturbação de 10^{-2} , por exemplo, a variável *X* apresentou redução de 80,51% em relação à média simples, enquanto *Y* apresentou 56,55% e *Z* apresentou 62,24%. Esses valores mostram que, no médio prazo, o agrupamento foi capaz de identificar representantes mais próximos da referência em cenários nos quais a média simples já não descrevia com a mesma precisão a estrutura do *ensemble*.

Ainda assim, esse comportamento não ocorreu da mesma forma em todas as configurações. Em cenários com perturbação inicial muito pequena, especialmente com 100 membros e perturbação de 10^{-4} , os percentuais tornaram-se negativos: -60,01% em *X*, -50,47% em *Y* e -52,18% em *Z*. Isso mostra que, mesmo no horizonte mais favorável ao agrupamento, a vantagem do método depende do grau de separação efetiva entre as trajetórias. Quando essa separação ainda é insuficiente, a média simples pode continuar sendo mais estável.

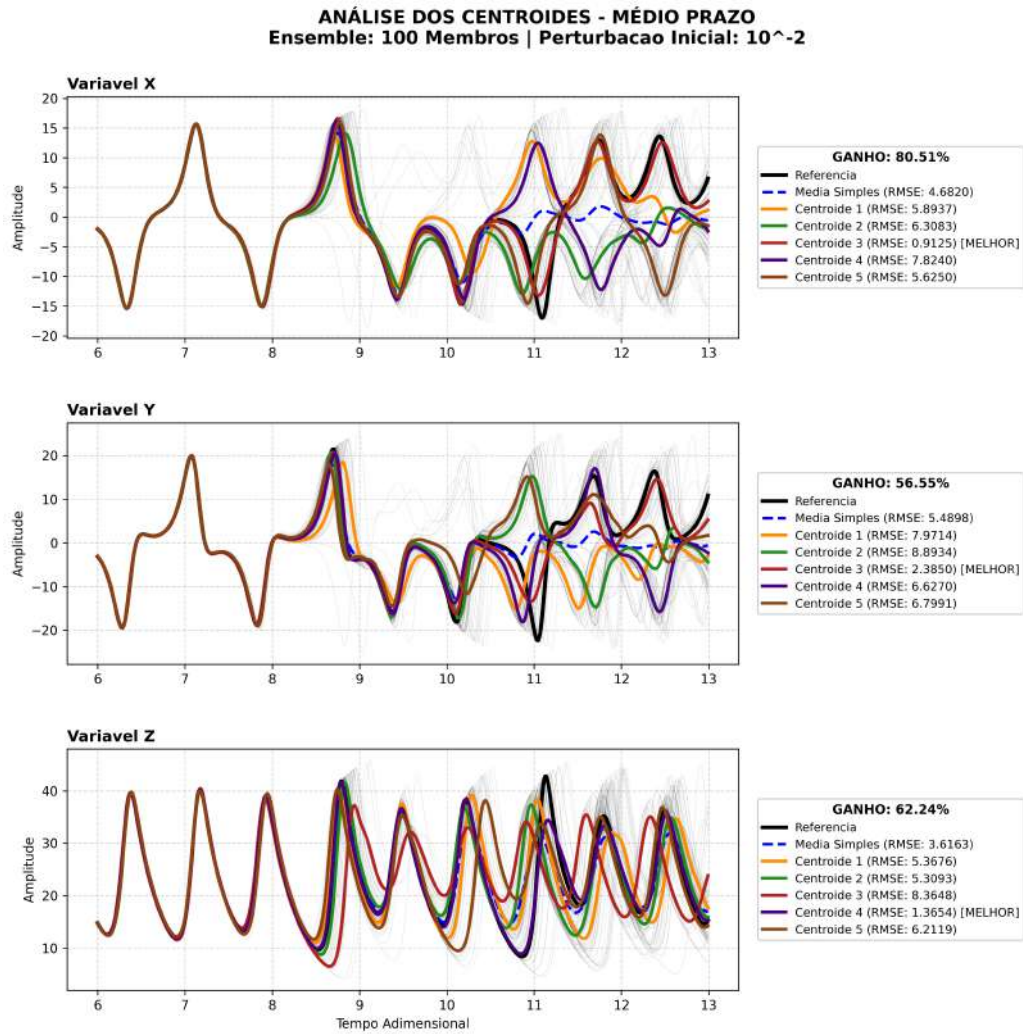


Figura 8 – Comportamento das variáveis X , Y e Z no médio prazo ($6 \leq t < 13$) para o cenário de 100 membros e ruído de 10^{-2} , com aplicação do K -means++. O gráfico mostra uma situação em que o agrupamento identifica uma trajetória mais próxima da referência do que a média aritmética simples.

Fonte: Autoria própria.

Tabela 3 – Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE no médio prazo ($6 \leq t < 13$), para 100 membros.

Ruído	Variável X	Variável Y	Variável Z
10^{-4}	-60,01%	-50,47%	-52,18%
10^{-3}	61,41%	46,00%	11,01%
10^{-2}	80,51%	56,55%	62,24%

Fonte: Autoria própria.

4.2.3 Longo Prazo ($13 \leq t < 20$)

Nos experimentos de longo prazo, os resultados indicam um comportamento menos regular para o agrupamento, condicionado pelo efeito de saturação da dispersão

discutido na Seção 4.1.3. Nessa fase, o sistema atinge um regime puramente caótico em que as trajetórias povoam o atrator de forma dispersa e homogênea. Como consequência, as diferenças métricas finas entre os estimadores se atenuam, tornando o comportamento estatístico flutuante e de menor relevância analítica para fins de prognóstico determinístico, uma vez que a variabilidade acumulada restringe a capacidade de ambos os estimadores em sustentar aderência à referência após a perda completa da memória das condições iniciais.

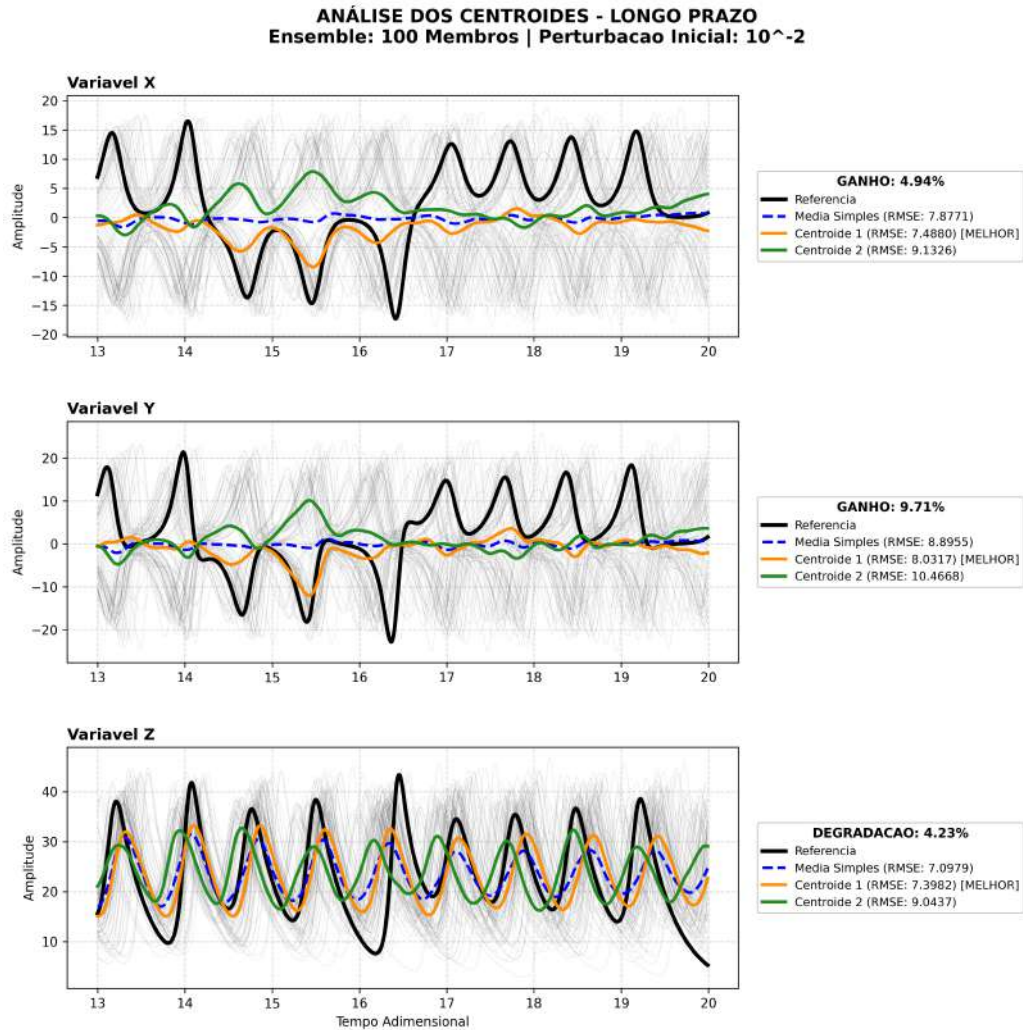


Figura 9 – Comportamento das variáveis X , Y e Z no longo prazo ($13 \leq t < 20$) para o cenário de 100 membros e perturbação de 10^{-2} , com aplicação do K -means++. O gráfico ilustra a saturação da dispersão em que a distribuição das trajetórias ocupa de forma homogênea os lóbulos do atrator estranho.

Fonte: Autoria própria.

4.3 Seleção do Número de Agrupamentos (K)

Nesta seção, analisam-se os valores de K determinados pelo Coeficiente de Silhueta. No algoritmo K -means++, o valor de K representa o número de grupos em que os dados são divididos, enquanto o centroide corresponde ao ponto representativo de

cada grupo formado. Para cada experimento, foram considerados os valores $K = 2, 3, 4$ e 5 , e a escolha final correspondeu ao valor que apresentou o maior coeficiente médio de silhueta. Esse critério foi adotado porque permite identificar, entre as opções testadas, a partição em que os grupos se mostram mais bem definidos, com maior proximidade entre os elementos de um mesmo grupo e maior separação em relação aos demais grupos (ROUSSEEUW, 1987). A Tabela 4 apresenta a distribuição da escolha de K para cada prazo analisado, servindo como base para as análises geométricas subsequentes.

Tabela 4 – Frequência de seleção de K por horizonte temporal.

Prazo	Valor de K mais frequente	Ocorrência	Comportamento dos Dados
Curto	$K = 3$	59,26%	Conjunto ainda coeso, com separações pouco profundas
Médio	$K = 2$	37,04%	Transição entre dois núcleos principais e partições mais refinadas
Longo	$K = 2$	55,56%	Dispersão acentuada, com maior heterogeneidade entre os casos

Fonte: Autoria própria.

No **curto prazo**, o valor mais frequente foi $K = 3$, presente em 16 dos 27 casos analisados, o que corresponde a 59,26% do total. Os demais 11 casos, equivalentes a 40,74%, apresentaram $K = 2$. Esse resultado mostra que, embora o *ensemble* ainda se encontre em um regime de baixa dispersão, o critério da Silhueta já identifica, com maior frequência, uma separação em três grupos. Ainda assim, a proximidade entre os membros permanece elevada, de modo que essa divisão reflete diferenças ainda sutis no interior do conjunto. Em outras palavras, a escolha de $K = 3$ não significa, necessariamente, uma fragmentação bem marcada do *ensemble*, mas apenas que, entre os valores testados, essa foi a opção que apresentou melhor separação estatística dos grupos nesse horizonte (WILKS, 2011).

No **médio prazo**, o valor mais frequente foi $K = 2$, com 10 ocorrências, correspondendo a 37,04% dos casos. Em seguida, observou-se $K = 5$ em 9 casos (33,33%), $K = 3$ em 6 casos (22,22%) e $K = 4$ em 2 casos (7,41%). Essa distribuição mostra que, neste horizonte, a escolha de K se torna mais variável. Em parte dos experimentos, a Silhueta identifica uma organização em dois grupos mais evidentes; em outros, aponta para partições mais refinadas. Isso indica que, nesta fase de transição, o *ensemble* pode assumir tanto uma configuração mais próxima de dois núcleos principais quanto uma estrutura mais fragmentada, dependendo da combinação entre número de membros, variável e intensidade de perturbação inicial (KALNAY, 2003; HASTIE *et al.*, 2009).

Já no **longo prazo**, o valor mais frequente voltou a ser $K = 2$, com 15 ocorrências, o que corresponde a 55,56% dos casos. Os demais resultados distribuíram-se entre

$K = 5$, com 5 ocorrências (18,52%), $K = 3$, com 4 ocorrências (14,81%), e $K = 4$, com 3 ocorrências (11,11%). Isso mostra que, mesmo sob maior dispersão, a escolha de dois grupos ainda foi a mais recorrente. No entanto, a presença dos demais valores de K indica que a configuração do *ensemble* se torna mais dependente das características específicas de cada experimento. Em alguns casos, a dispersão acumulada ainda permite uma organização em dois grupos mais amplos; em outros, o algoritmo seleciona partições mais refinadas para descrever a distribuição dos membros. Desse modo, a escolha de K nessa escala temporal passa a refletir a maior variabilidade geométrica do conjunto (JAIN, 2010; YUSSOUF *et al.*, 2004).

4.4 Análise da Distribuição Espacial dos Agrupamentos

A investigação da distribuição espacial dos membros no espaço de fases permite compreender com maior clareza os resultados obtidos com a técnica de agrupamento. Enquanto os gráficos temporais mostram a evolução das trajetórias ao longo de cada horizonte, as projeções bidimensionais permitem observar, no instante final de cada intervalo, como os membros se distribuem no espaço de fases. Essa visualização torna mais evidente o grau de dispersão do conjunto, a formação de regiões de maior concentração e a posição relativa da média aritmética simples, dos centróides e da trajetória de referência.

Essa leitura é particularmente importante para interpretar, de forma mais clara, por que a média aritmética simples ou o centróide selecionado permanecem mais próximos da trajetória de referência em determinados cenários. Quando os membros do *ensemble* se distribuem de forma mais concentrada, a média tende a representar adequadamente o conjunto. Quando passam a ocupar regiões espacialmente distintas, o agrupamento pode identificar centróides mais representativos da organização do *ensemble* naquele instante.

4.4.1 Distribuição Espacial no Curto Prazo

No curto prazo, a visualização bidimensional mostra que os membros do *ensemble* ainda permanecem fortemente concentrados no espaço de fases. No instante final desse horizonte, a nuvem de pontos apresenta pequena dispersão, e a trajetória de referência permanece próxima da região central ocupada pelo conjunto. Nessas condições, a média aritmética simples continua sendo uma representação estável, pois a concentração dos membros ainda não produz separações espaciais suficientemente marcadas.

No cenário com 100 membros e perturbação de 10^{-2} , essa configuração ajuda a compreender por que o agrupamento não apresentou vantagem. Como a distribuição espacial do *ensemble* ainda se mantém bastante compacta, os centróides calculados

passam a representar subdivisões internas de uma nuvem pouco dispersa. Nesse caso, a média simples permanece mais próxima da região central do conjunto e, conseqüentemente, mais próxima da trajetória de referência.

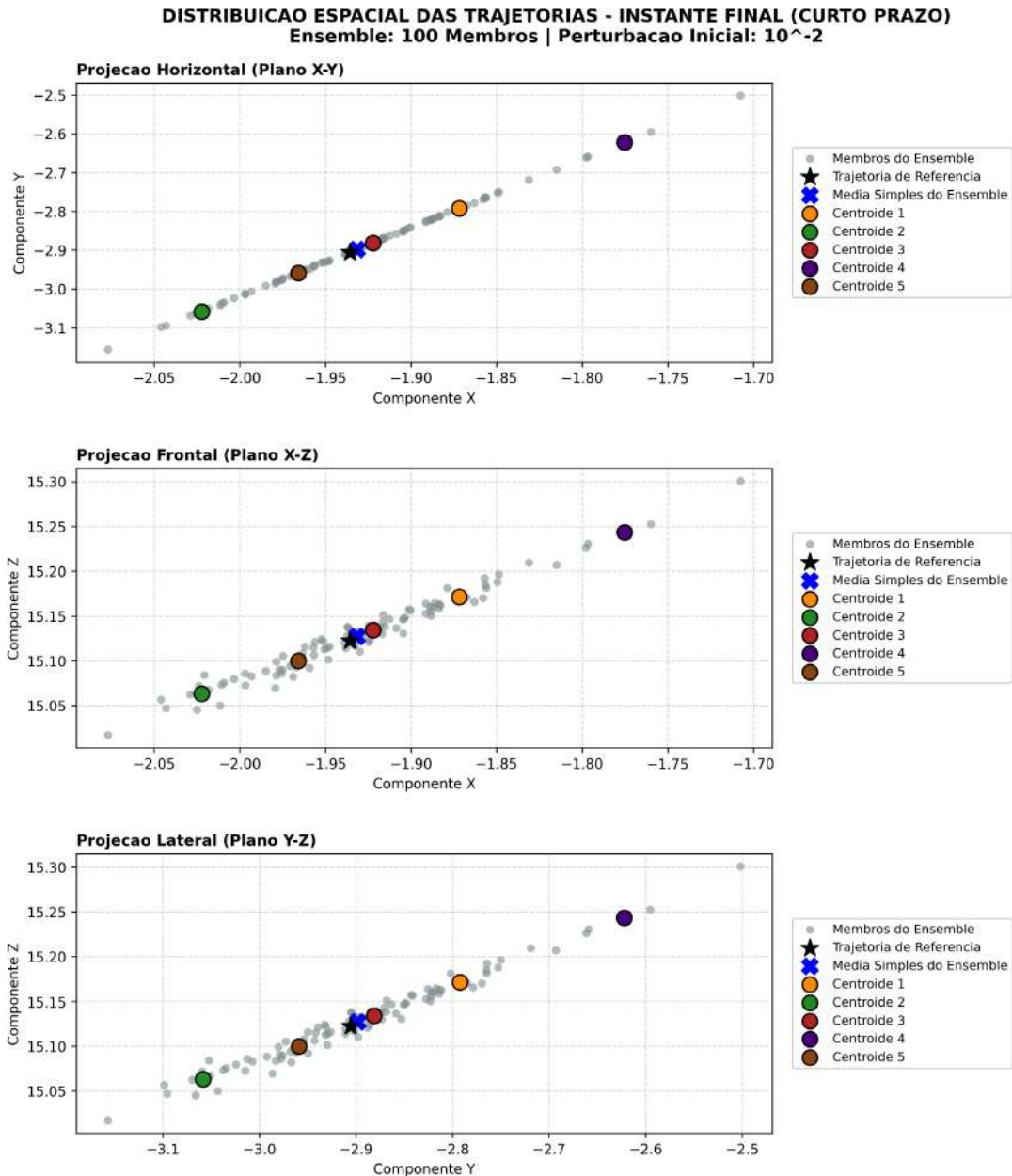


Figura 10 – Distribuição espacial das variáveis X , Y e Z no curto prazo ($0 \leq t < 6$) para o cenário com 100 membros e ruído de 10^{-2} . A concentração dos membros no instante final do intervalo evidencia a baixa dispersão do conjunto e a proximidade entre a média aritmética simples e a trajetória de referência.

Fonte: Autoria própria.

4.4.2 Distribuição Espacial no Médio Prazo

No médio prazo, a representação bidimensional mostra com maior clareza o aumento da dispersão do *ensemble* e a formação de regiões espacialmente mais diferenciadas. No instante final desse horizonte, os membros já não permanecem con-

centrados em uma única região, e a distribuição do conjunto passa a revelar estruturas mais separadas no espaço de fases.

Nessa situação, a média aritmética simples tende a ocupar uma posição intermediária em relação à nuvem de pontos. Quando a trajetória de referência se encontra próxima de uma dessas regiões mais concentradas, o centróide associado a esse agrupamento pode permanecer mais próximo da referência do que a média simples. Assim, a visualização espacial no instante final do médio prazo evidencia de forma mais direta por que o agrupamento passa a ser vantajoso: a referência deixa de estar associada ao centro geométrico do conjunto e passa a se localizar mais próxima de uma região específica da distribuição, melhor representada pelo centróide correspondente, mitigando o viés de suavização linear da média convencional.

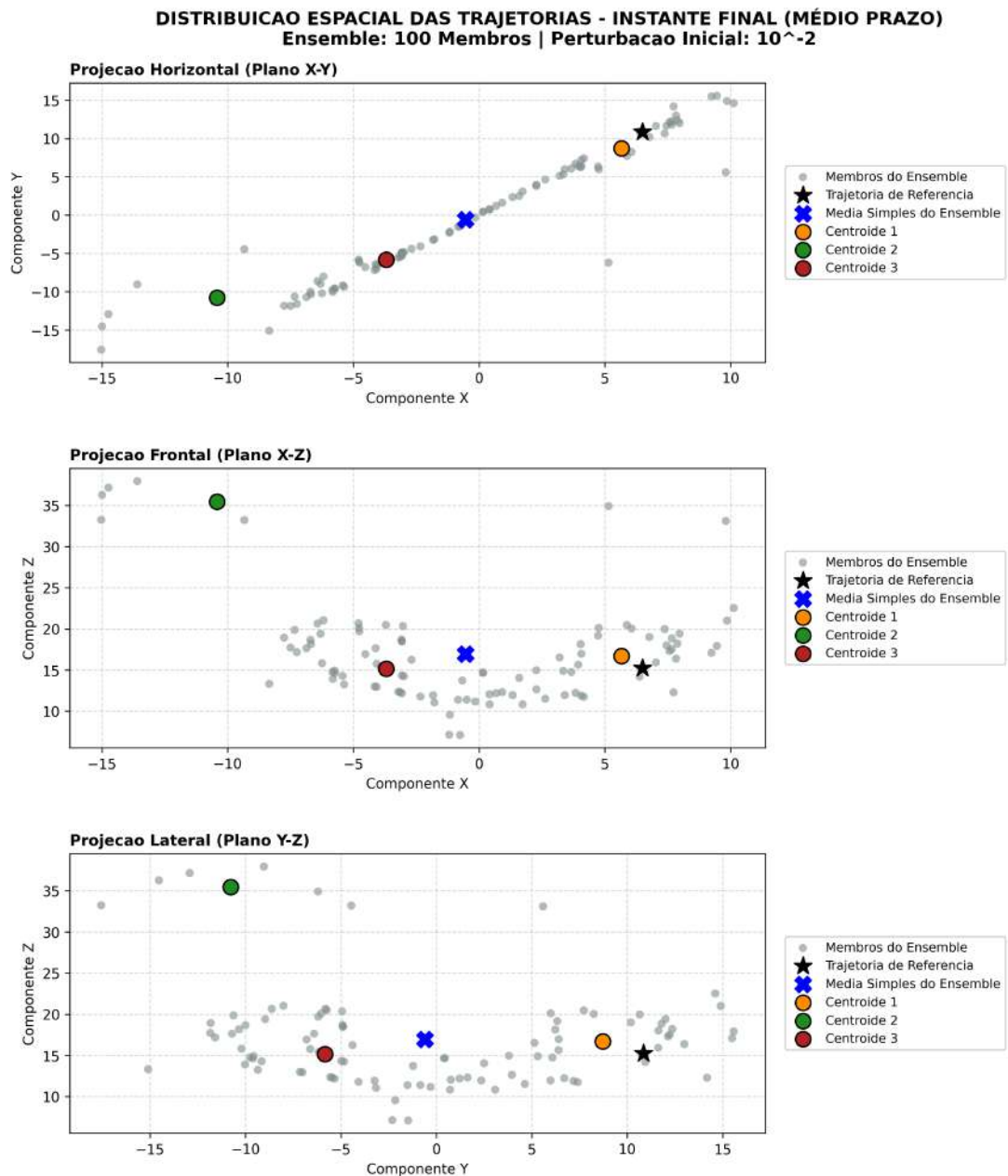


Figura 11 – Distribuição espacial das variáveis X , Y e Z no médio prazo ($6 \leq t < 13$) para o cenário com 100 membros e ruído de 10^{-2} . A configuração no instante final mostra maior separação entre regiões do conjunto, favorecendo a proximidade entre o centróide selecionado e a trajetória de referência.

Fonte: Autoria própria.

4.4.3 Distribuição Espacial no Longo Prazo

No longo prazo, a representação bidimensional atesta a consolidação da saturação espacial do *ensemble*. Os membros dispersam-se amplamente sobre a área total das estruturas atratoras, tornando o conjunto altamente heterogêneo. Conforme postulado na subseção 4.2.3, as assinaturas geométricas tornam-se difusas nesta escala temporal, o que atenua as diferenças práticas entre os centroides e o operador de média linear frente à solução de referência do modelo de Lorenz.

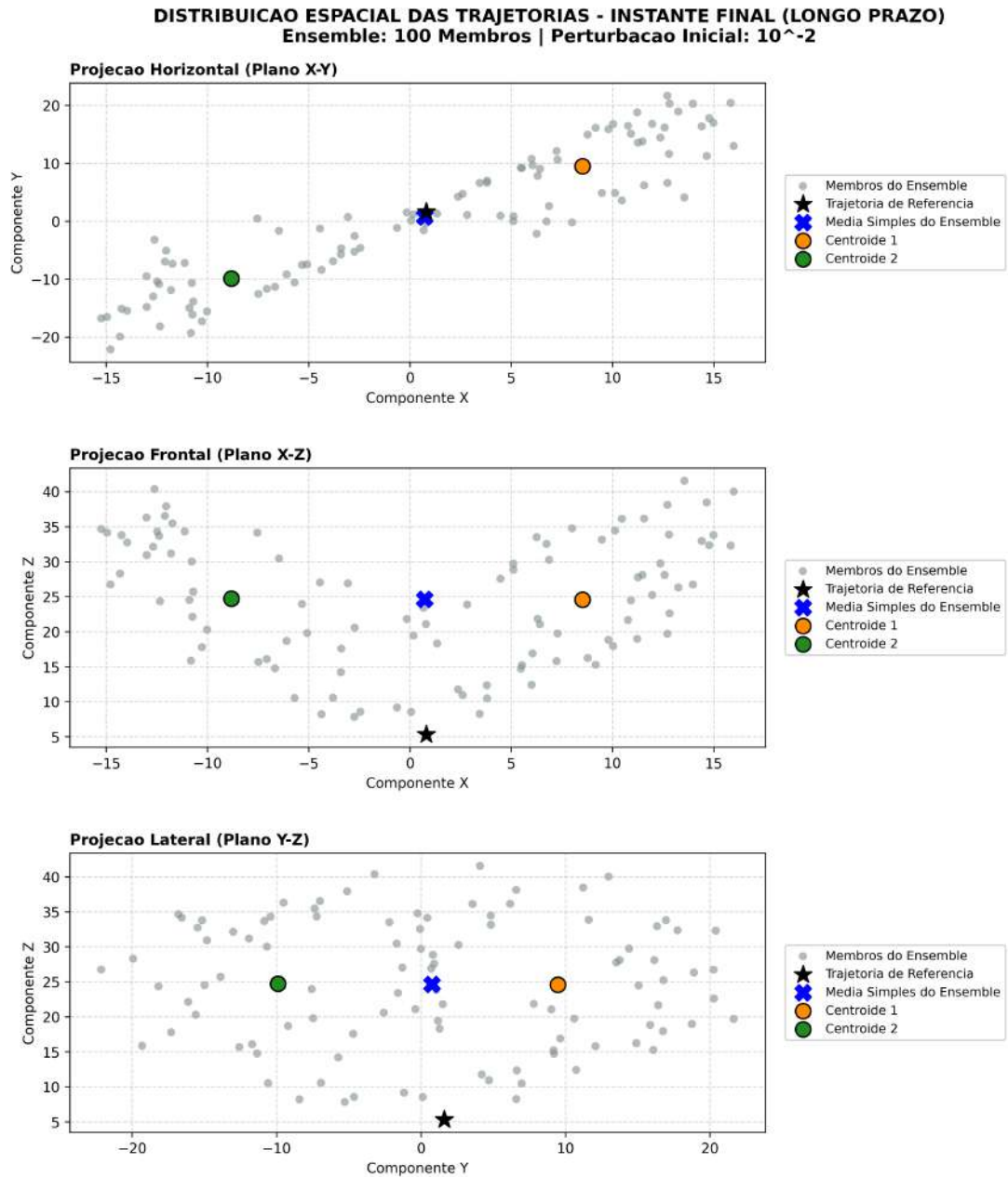


Figura 12 – Distribuição espacial das variáveis X , Y e Z no longo prazo ($13 \leq t < 20$) para o cenário com 100 membros e ruído de 10^{-2} . A maior dispersão dos membros no instante final reduz o contraste entre a média aritmética simples e os centróides como representantes da trajetória de referência.

Fonte: Autoria própria.

4.5 Análise Comparativa das Variáveis X , Y e Z

Nesta seção, analisa-se o comportamento das variáveis X , Y e Z em relação à técnica de agrupamento, considerando o conjunto completo dos cenários avaliados. O objetivo é identificar em que medida o agrupamento superou a média aritmética simples em cada componente do sistema e como essas diferenças podem ser interpretadas à luz do comportamento observado ao longo dos horizontes temporais.

Para cada variável, foram considerados 27 casos, correspondentes às combinações entre número de membros do *ensemble*, intensidade de perturbação nas condições iniciais e horizonte temporal. Na Tabela 5, apresenta-se a frequência de casos em que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples.

Tabela 5 – Distribuição dos casos em que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples, por variável.

Variável	Casos	Percentual	Síntese
X	19 de 27	70,37%	Maior frequência de casos em que o agrupamento superou a média simples
Y	19 de 27	70,37%	Comportamento semelhante ao observado em X
Z	17 de 27	62,96%	Menor frequência relativa de casos em que o agrupamento superou a média simples

Fonte: Autoria própria.

Os resultados mostram que o agrupamento apresentou comportamento competitivo nas três variáveis analisadas, embora com diferenças na frequência de casos em que superou a média simples. As variáveis X e Y concentraram 19 casos em 27 nos quais o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples, correspondendo a 70,37% do total em cada uma. A variável Z , por sua vez, apresentou 17 casos em 27, o que corresponde a 62,96%. Esses valores indicam que o agrupamento superou a média simples com maior frequência em X e Y , embora também tenha apresentado esse comportamento em parte significativa dos casos envolvendo Z .

4.5.1 Análise das Variáveis X e Y

As variáveis X e Y apresentaram o maior número de casos em que o centróide de menor RMSE superou a média aritmética simples. Esse resultado sugere que, nas componentes em que a separação entre trajetórias se torna mais evidente, o agrupamento tende a oferecer maior vantagem em relação à média simples.

No médio prazo, essa diferença torna-se mais clara. Nessa fase, o *ensemble* deixa de permanecer concentrado em uma única região, e a distribuição dos membros passa a apresentar separações mais visíveis. Nas componentes de estado X e Y , o comportamento caótico bimodal impera sobre a evolução dinâmica, forçando os membros a se dividirem entre os polos simétricos e opostos do atrator estranho. Nessa situação de severa divergência espacial, a média aritmética simples tende a ocupar uma posição intermediária e vazia em relação ao conjunto, enquanto o agrupamento consegue identificar com fidelidade a trajetória física de um dos subgrupos funcionais. Isso ajuda a explicar por que X e Y apresentaram maior frequência de casos em que o centróide de menor RMSE ficou mais próximo da referência do que a média simples.

Entre as duas variáveis, X e Y mostraram comportamento bastante próximo em

termos de frequência total. Ainda assim, os resultados anteriores indicam que X tende a apresentar, em alguns cenários, diferenças mais acentuadas no médio prazo. Isso sugere que a organização do *ensemble* nessa componente favorece, com maior nitidez, a identificação de grupos representativos em situações de maior separação entre trajetórias.

4.5.2 Análise da Variável Z

A variável Z também apresentou casos em que o agrupamento superou a média simples, mas com frequência menor do que X e Y . O total de 17 casos em 27 mostra que o centróide de menor RMSE apresentou valor inferior ao da média aritmética simples em parte dos cenários, embora isso tenha ocorrido com menor intensidade do que nas outras duas variáveis.

Esse comportamento diferenciado é explicado diretamente pelas propriedades geométricas e físicas intrínsecas da equação que governa a coordenada Z . Enquanto X e Y controlam a alternância e a inversão polar entre as duas asas do atrator de Lorenz, a variável Z quantifica o desvio do perfil vertical de temperatura e permanece estritamente positiva ($Z > 0$) durante todo o regime de integração fluida. Por não experimentar inversões de sinal, o comportamento caótico bimodal não impera sobre a variável Z ; em vez disso, suas trajetórias estruturam-se em órbitas concêntricas ao redor de um eixo fixo, preservando uma geometria de dispersão intrinsecamente unimodal. Devido a essa distribuição espacial compacta, a média aritmética simples consegue reter uma robusta aderência ao centro de massa das trajetórias por intervalos mais longos, tornando o agrupamento estatisticamente menos vantajoso ou redundante para essa coordenada específica em vários cenários.

Ainda assim, no médio prazo, a variável Z também apresentou casos em que o agrupamento se mostrou superior à média simples, sobretudo quando a magnitude da dispersão do *ensemble* expandiu o diâmetro das órbitas o suficiente para distorcer a linearidade central da média. Isso mostra que a variável Z não é incompatível com a técnica de agrupamento, mas responde a ela de forma menos intensa do que X e Y no conjunto dos cenários analisados.

4.5.3 Comparação entre as Variáveis ao Longo do Tempo

A análise conjunta das três variáveis mostra que a vantagem do agrupamento não se distribui de forma idêntica entre as componentes do sistema. No curto prazo, quando o *ensemble* ainda se encontra mais concentrado, a diferença entre média simples e agrupamento tende a ser menos marcada. No médio prazo, com o aumento da dispersão e a separação entre trajetórias, o agrupamento passa a superar a média simples com maior frequência, sobretudo em X e Y . No longo prazo, sob o advento da saturação caótica generalizada, essa vantagem torna-se irregular e secundária para

todas as variáveis, dada a homogeneização da dispersão no atrator.

Desse modo, a comparação por variável reforça a interpretação já observada nas seções anteriores: o desempenho do agrupamento depende da forma assumida pelo *ensemble* em cada horizonte temporal. As variáveis X e Y mostraram-se mais sensíveis às situações em que a separação bimodal entre trajetórias favorece a identificação de grupos mais representativos. A variável Z , devido à sua dinâmica concêntrica e unimodal, mostrou comportamento relativamente mais próximo ao da média no conjunto dos experimentos.

De modo geral, os resultados indicam que o agrupamento superou a média simples com maior frequência nas variáveis X e Y , sem deixar de ser relevante para Z . Essa diferença entre as componentes contribui para compreender de forma mais completa em quais situações o *K-means++* pode oferecer uma representação mais adequada do *ensemble* do que a média aritmética simples.

4.6 Análise Comparativa dos Níveis de Ruído

A magnitude da perturbação inicial influencia diretamente a velocidade com que as trajetórias do *ensemble* se separam ao longo do tempo. Quando a incerteza inicial é menor, os membros tendem a permanecer mais próximos entre si por mais tempo. À medida que a perturbação nas condições iniciais aumenta, a dispersão entre as trajetórias se torna mais evidente, alterando a relação entre a média aritmética simples, os centróides e a trajetória de referência.

No cenário com perturbação de 10^{-4} , a separação entre as trajetórias ainda é limitada no médio prazo. Nessa condição, a média aritmética simples permanece próxima da estrutura principal do conjunto, o que ajuda a explicar os resultados negativos observados nesse horizonte para 100 membros. A distribuição bidimensional mostra que, nesse caso, o *ensemble* ainda não apresenta separação suficientemente marcada para que a análise geométrica por agrupamento produza uma representação mais próxima da referência.

Quando a perturbação inicial é elevada para 10^{-3} , observa-se uma mudança mais nítida na configuração do *ensemble* no médio prazo. A dispersão passa a revelar regiões mais separadas no espaço de fases, favorecendo a identificação de agrupamentos mais consistentes. Nesse cenário, o centróide selecionado passa a representar melhor uma dessas regiões do que a média simples, o que explica a mudança observada nos resultados.

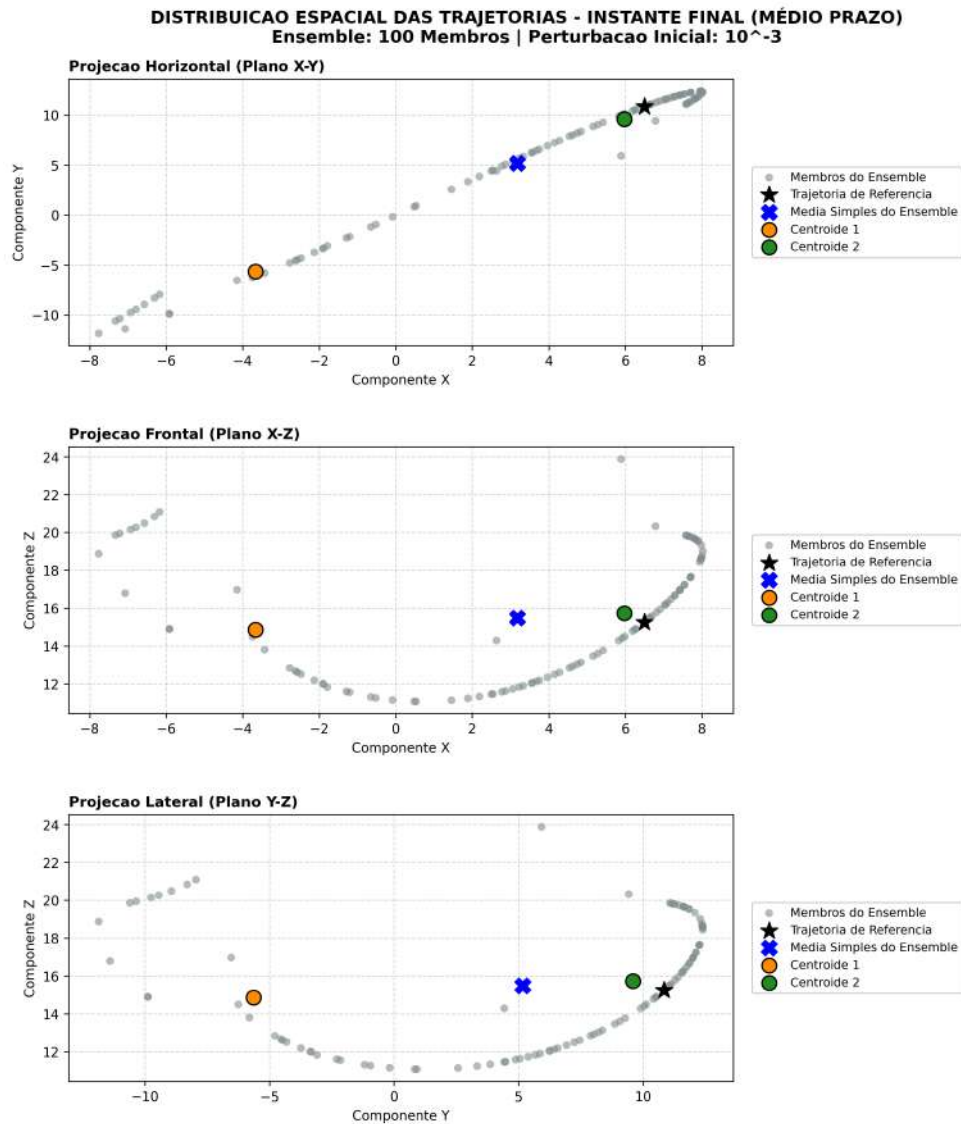


Figura 13 – Distribuição espacial no médio prazo ($6 \leq t < 13$) para o cenário com 100 membros e ruído de 10^{-3} . A maior separação entre os membros favorece a identificação de grupos mais definidos.

Fonte: Autoria própria.

Sob perturbação de 10^{-2} , a separação entre as trajetórias ocorre de forma mais célere. Nesse caso, a organização espacial do *ensemble* favorece de maneira mais clara a identificação de regiões distintas, permitindo que o agrupamento selecione centróides mais próximos da trajetória de referência, especialmente no médio prazo.

Tabela 6 – Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE por nível de ruído, para a variável X com 100 membros.

Nível de Ruído	Curto Prazo	Médio Prazo	Longo Prazo
10^{-4}	-317,77%	-60,01%	24,02%
10^{-3}	-327,34%	61,41%	8,43%
10^{-2}	-309,03%	80,51%	4,94%

Fonte: Autoria própria.

A Tabela 6 mostra que o efeito da incerteza inicial varia conforme o horizonte temporal. No curto prazo, os percentuais permanecem negativos nos três níveis de perturbação para o cenário com 100 membros, indicando que a média simples ainda se manteve mais próxima da trajetória de referência em virtude da baixa dispersão inicial. No médio prazo, janela de dispersão intermediária, os resultados passam a ser positivos para 10^{-3} e 10^{-2} , o que atesta que o aumento da perturbação nas condições iniciais acelerou a separação bimodal e expandiu a utilidade metodológica do agrupamento. No longo prazo, sob saturação, os valores aproximam-se do limiar de atenuação métrica.

4.6.1 Sensibilidade à Densidade do Ensemble

A resposta do agrupamento também depende do número de membros utilizados em cada experimento. À medida que o tamanho do *ensemble* aumenta, a distribuição das trajetórias no espaço de fases pode ser observada com maior detalhe. No entanto, os resultados mostram que esse efeito não é uniforme em todos os horizontes temporais.

Com 15 membros, o agrupamento já apresentou desempenho positivo no médio prazo. Isso mostra que, mesmo em conjuntos menores, a técnica pode identificar uma região representativa do *ensemble* quando a separação entre as trajetórias já está suficientemente definida. Ainda assim, a menor quantidade de pontos torna a posição dos centróides mais sensível à distribuição específica dos membros.

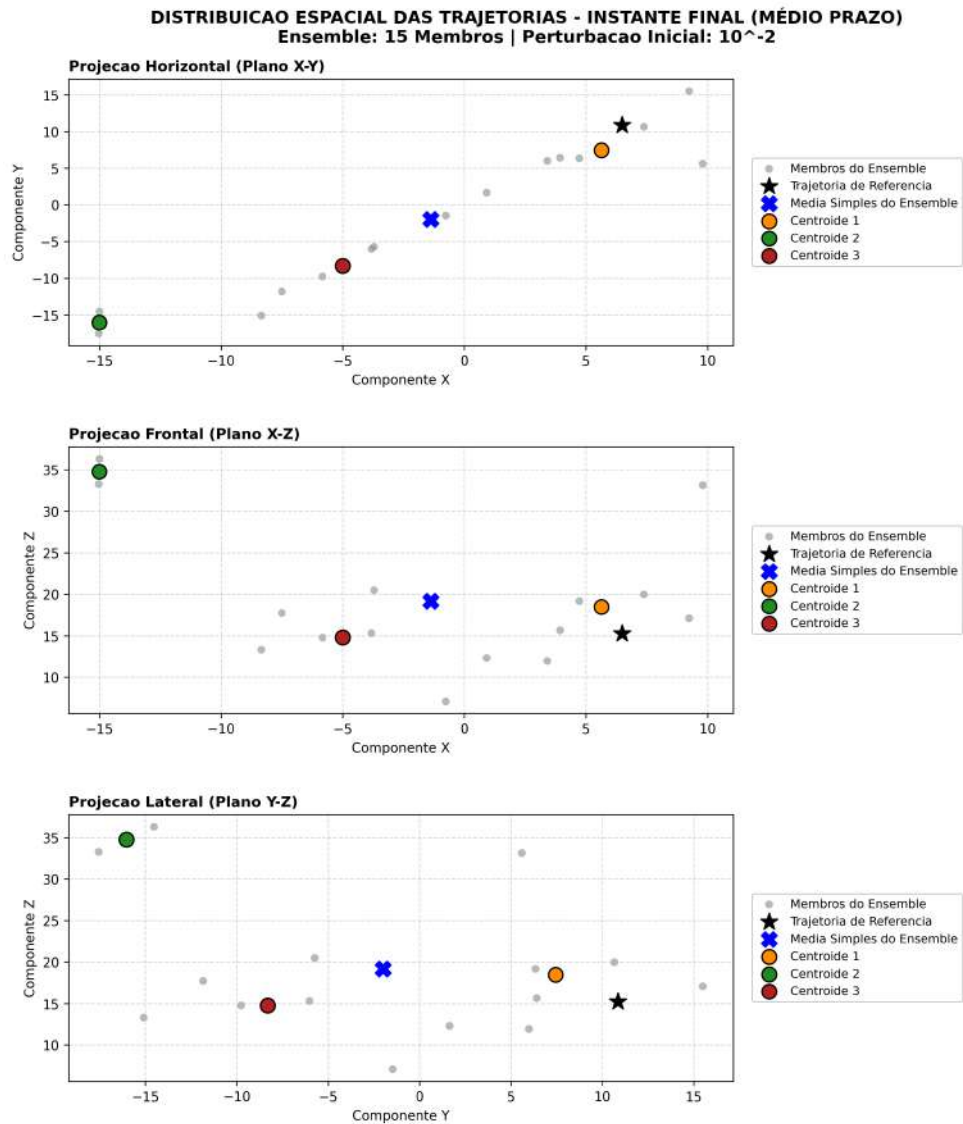


Figura 14 – Distribuição espacial no médio prazo ($6 \leq t < 13$) para o cenário com 15 membros e ruído de 10^{-2} . A menor quantidade de pontos reduz o detalhamento da distribuição do conjunto.

Fonte: Autoria própria.

Com 50 membros, a distribuição espacial do *ensemble* passa a ser visualizada de forma mais estável. Nesse caso, a organização dos membros nas regiões mais concentradas do espaço de fases torna-se mais nítida, favorecendo a identificação de grupos representativos e contribuindo para a melhora dos resultados observados no médio prazo.

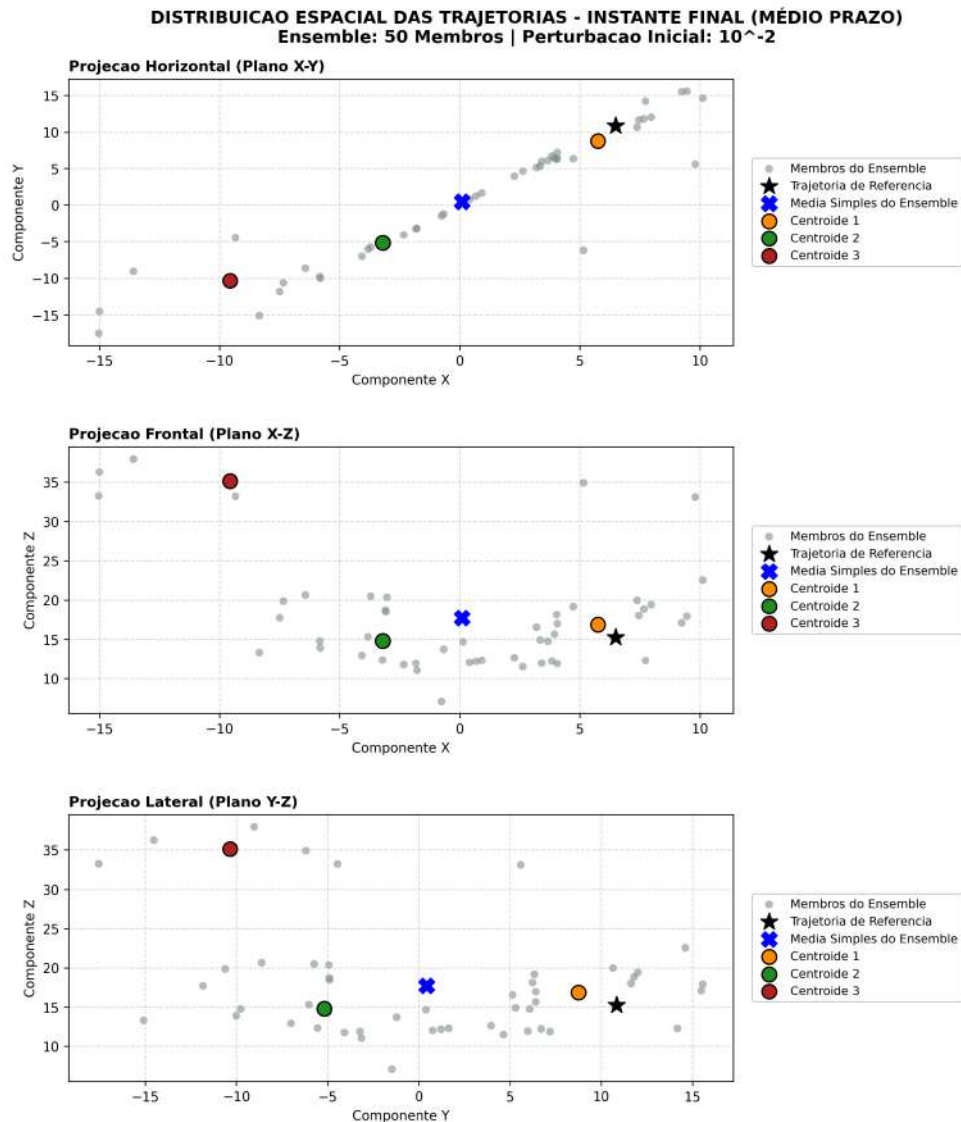


Figura 15 – Distribuição espacial no médio prazo ($6 \leq t < 13$) para o cenário com 50 membros e ruído de 10^{-2} . O aumento do número de membros permite visualizar com maior clareza a organização do conjunto.

Fonte: Autoria própria.

No caso de 100 membros, a distribuição do *ensemble* é ainda mais detalhada. Entretanto, os resultados mostram que o aumento no número de membros não implica melhora automática em todos os horizontes. No curto prazo, por exemplo, os percentuais sofreram a instabilidade de escala discutida na seção 4.2.1, enquanto no médio prazo observou-se o maior valor positivo e de maior previsibilidade real entre os cenários analisados. Isso indica que o efeito da densidade do *ensemble* depende do cruzamento entre o número total de membros e o estágio de maturação da dispersão espacial.

Tabela 7 – Diferença relativa percentual entre o RMSE da média aritmética simples e o RMSE do centróide de menor RMSE por número de membros, para a variável X com ruído de 10^{-2} .

Número de Membros	Curto Prazo	Médio Prazo	Longo Prazo
15 membros	61,91%	60,87%	-13,23%
50 membros	76,75%	66,85%	16,26%
100 membros	-309,03%	80,51%	4,94%

Fonte: Autoria própria.

A Tabela 7 resume que a eficiência prática da abordagem reside primariamente na aplicabilidade do diagnóstico de subgrupos em horizontes de médio prazo. A identificação de partições estruturadas fornece uma ferramenta matemática robusta para subsidiar sistemas de tomada de decisão, mapeando cenários alternativos divergentes que seriam suprimidos pela suavização linear da média.

Cabe ressaltar que, devido à grande quantidade e extensão dos cenários simulados combinando diferentes tamanhos de conjunto, horizontes temporais e intensidades de perturbação nas condições iniciais, todos os gráficos estruturais e tabelas estatísticas remanescentes que complementam e ratificam as análises discutidas encontram-se catalogados para consulta nos Apêndices deste trabalho.

5 CONCLUSÕES

A análise estatística dos experimentos realizados demonstrou que o algoritmo *K-means++* cumpre uma importante função metodológica na organização e síntese de previsões por conjuntos em sistemas caóticos de baixa dimensão. No panorama global dos cenários avaliados, o centróide de menor RMSE apresentou erro inferior ao da média aritmética simples em 67,90% dos casos, confirmando que a identificação de trajetórias representativas extraídas da estrutura geométrica do conjunto é uma alternativa viável aos métodos tradicionais de agregação.

Os resultados revelaram que a eficácia do agrupamento é estritamente condicionada pelo horizonte temporal e pelo nível de dispersão do *ensemble*. O método encontrou sua janela ótima de aplicação no médio prazo, horizonte no qual superou a média simples em 88,89% dos casos. Nessa fase intermediária, a separação induzida pelas perturbações iniciais (especialmente nas magnitudes de 10^{-3} e 10^{-2}) é suficiente para que o algoritmo separe trajetórias com evoluções dinâmicas distintas, capturando regimes de comportamento que a média simples, por ser propensa a estados suavizados em regiões de baixa probabilidade, não consegue representar.

Em contrapartida, nos extremos temporais do sistema, a vantagem comparativa do agrupamento se atenua devido a restrições dinâmicas intrínsecas. No curto prazo, a forte aproximação inicial das trajetórias mantém os valores absolutos de erro reduzidos, o que valida a suficiência da média simples para a maioria das configurações. No longo prazo, o efeito de saturação da dispersão faz com que os membros ocupem o atrator estranho de forma homogênea; a perda completa da memória das condições iniciais resulta em um comportamento estatístico flutuante, limitando a capacidade de ambos os estimadores em sustentar aderência à solução de referência.

Dessa forma, conclui-se que a principal contribuição desta dissertação não reside na substituição sistemática da média simples, mas na demonstração de que o agrupamento individualizado por variável identifica representantes mais aderentes à referência sob condições específicas de espalhamento do conjunto. Como limitação, ressalta-se o caráter exploratório em um modelo de baixa dimensão a partir de uma simulação aleatória controlada. Como desdobramentos e trabalhos futuros, sugere-se

a expansão desta metodologia para análises multivariadas integradas e a aplicação do método em modelos numéricos operacionais de previsão do tempo com maior realismo físico.

REFERÊNCIAS

- ALHAMED, A. *et al.* Cluster analysis of multimodel ensemble data from SAMEX. **Monthly Weather Review**, v. 130, n. 2, p. 226–256, 2002.
- ARTHUR, D. *et al.* k-means++: The advantages of careful seeding. In: **Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms**. New Orleans: Society for Industrial and Applied Mathematics, 2007. p. 1027–1035.
- BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.
- BOYCE, W. E. *et al.* **Equações Diferenciais Elementares e Problemas de Valores de Contorno**. 12. ed. Rio de Janeiro: LTC, 2024.
- BUIZZA, R. *et al.* A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. **Monthly Weather Review**, v. 133, p. 1076–1097, 2005.
- BUIZZA, R. *et al.* The singular-vector structure of the atmospheric global circulation. **Journal of the Atmospheric Sciences**, v. 52, p. 1434–1456, 1995.
- CALVETTI, L. **Previsão hidrometeorológica probabilística na Bacia do Alto Iguaçu-PR com os modelos WRF e TopModel**. Tese (Tese (Doutorado em Meteorologia)) — Universidade de São Paulo, São Paulo, 2011.
- EPSTEIN, E. S. Stochastic dynamic prediction. **Tellus**, v. 21, n. 6, p. 739–759, 1969.
- FORGY, E. W. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. **Biometrics**, v. 21, n. 3, p. 768–769, 1965. Abstract.
- GHIL, M. *et al.* **Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics**. New York: Springer-Verlag, 1987.
- HASTIE, T. *et al.* **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York: Springer, 2009.
- HINDMARSH, A. C. Odepack, a systematized collection of ode solvers. In: **Scientific Computing**. Amsterdam: North-Holland, 1983. p. 55–64. R. S. Stepleman *et al.* (ed.).
- HUANG, J. *et al.* The seasonality of predictability in the Lorenz model. **Journal of the Atmospheric Sciences**, v. 53, p. 2046–2063, 1996.
- JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, v. 31, p. 651–666, 2010.

JENNRICH, G. *et al.* Ensemble predictability of week 3/4 precipitation and temperature over the united states via cluster analysis of the large-scale circulation. **Weather and Forecasting**, v. 39, p. 1531–1544, 2024.

JOHNSON, C. *et al.* On the reliability and calibration of ensemble forecasts. **Monthly Weather Review**, v. 137, n. 5, p. 1717–1720, 2009.

KALNAY, E. **Atmospheric Modeling, Data Assimilation and Predictability**. Cambridge: Cambridge University Press, 2003.

KALNAY, E. *et al.* The NCEP/NCAR 40-year reanalysis project. **Bulletin of the American Meteorological Society**, v. 77, p. 437–472, 1996.

KIEL, B. M. *et al.* Comparison of clustering approaches in a multimodel ensemble for U.S. east coast cold season extratropical cyclones. **Weather and Forecasting**, v. 39, n. 3, p. 461–484, 2024.

KRISHNAMURTI, T. N. *et al.* Improved weather and seasonal climate forecasts from multimodel superensemble. **Science**, v. 285, p. 1548–1550, 1999.

LAMBERSON, W. S. *et al.* The use of ensemble clustering on a multimodel ensemble for medium-range forecasting at the weather prediction center. **Weather and Forecasting**, v. 38, n. 4, p. 539–554, 2023.

LEITH, C. E. Theoretical skill of Monte Carlo forecasting. **Monthly Weather Review**, v. 102, n. 6, p. 409–418, 1974.

LEUTBECHER, M. *et al.* Ensemble forecasting. **Journal of Computational Physics**, v. 227, n. 7, p. 3515–3539, 2008.

LLOYD, S. P. Least squares quantization in PCM. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129–137, 1982.

LORENZ, E. N. Deterministic nonperiodic flow. **Journal of the Atmospheric Sciences**, v. 20, n. 2, p. 130–141, 1963.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**. Berkeley: University of California Press, 1967. v. 1, p. 281–297.

PETZOLD, L. R. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. **SIAM Journal on Scientific and Statistical Computing**, v. 4, n. 1, p. 136–148, 1983.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987.

SPARROW, C. **The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors**. New York: Springer-Verlag, 1982.

STENSRUD, D. J. *et al.* Using initial condition and model physics perturbations in short-range ensemble forecasting of mesoscale convective systems. **Monthly Weather Review**, v. 128, p. 2077–2107, 2000.

THORNDIKE, R. L. Who belongs in the family? **Psychometrika**, v. 18, n. 4, p. 267–276, 1953.

TOTH, Z. *et al.* Ensemble forecasting at NMC: The generation of perturbations. **Bulletin of the American Meteorological Society**, v. 74, p. 2317–2330, 1993.

WILKS, D. S. **Statistical Methods in the Atmospheric Sciences**. 2nd. ed. Oxford: Academic Press, 2005.

WILKS, D. S. **Statistical Methods in the Atmospheric Sciences**. 3rd. ed. Oxford: Academic Press, 2011.

YUSSOUF, N. *et al.* Cluster analysis of multimodel ensemble data over New England. **Monthly Weather Review**, v. 132, p. 2452–2466, 2004.

APÊNDICES

APÊNDICE A – Análise gráfica e tabelas complementares do horizonte temporal de curto prazo

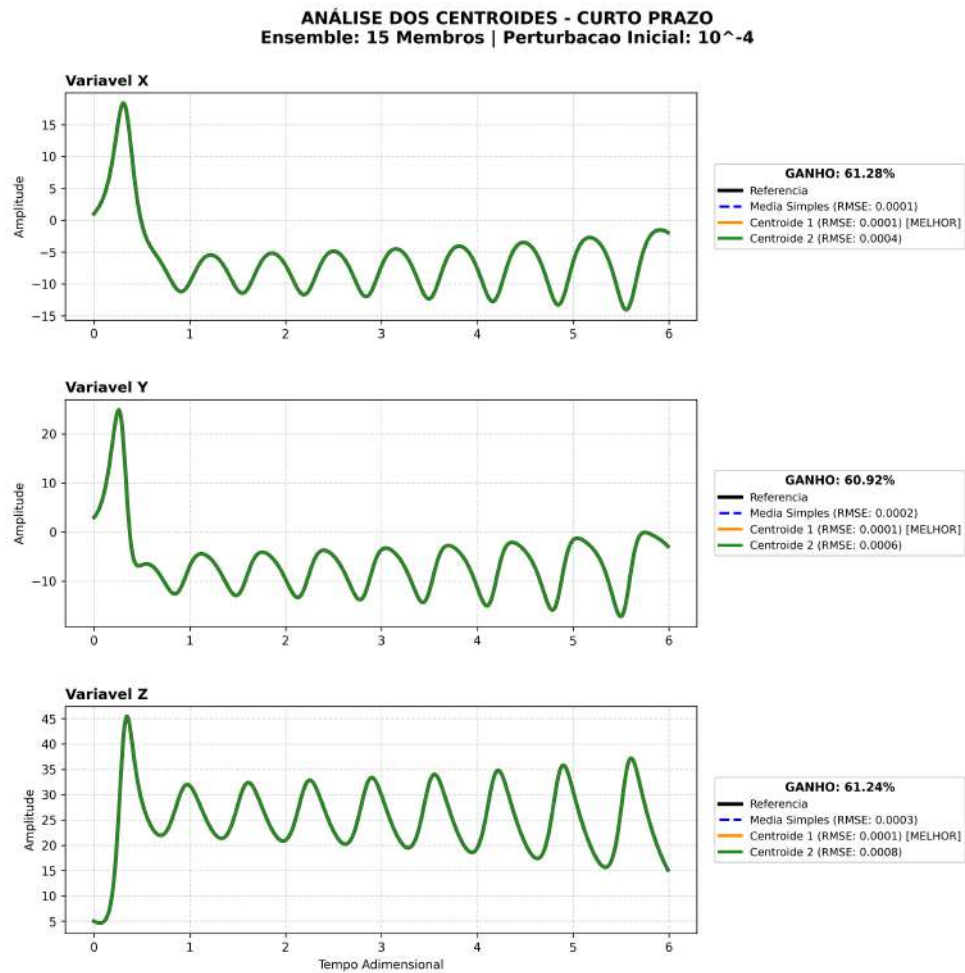


Figura 16 – Curto prazo ($0 \leq t < 6$), $N = 15$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

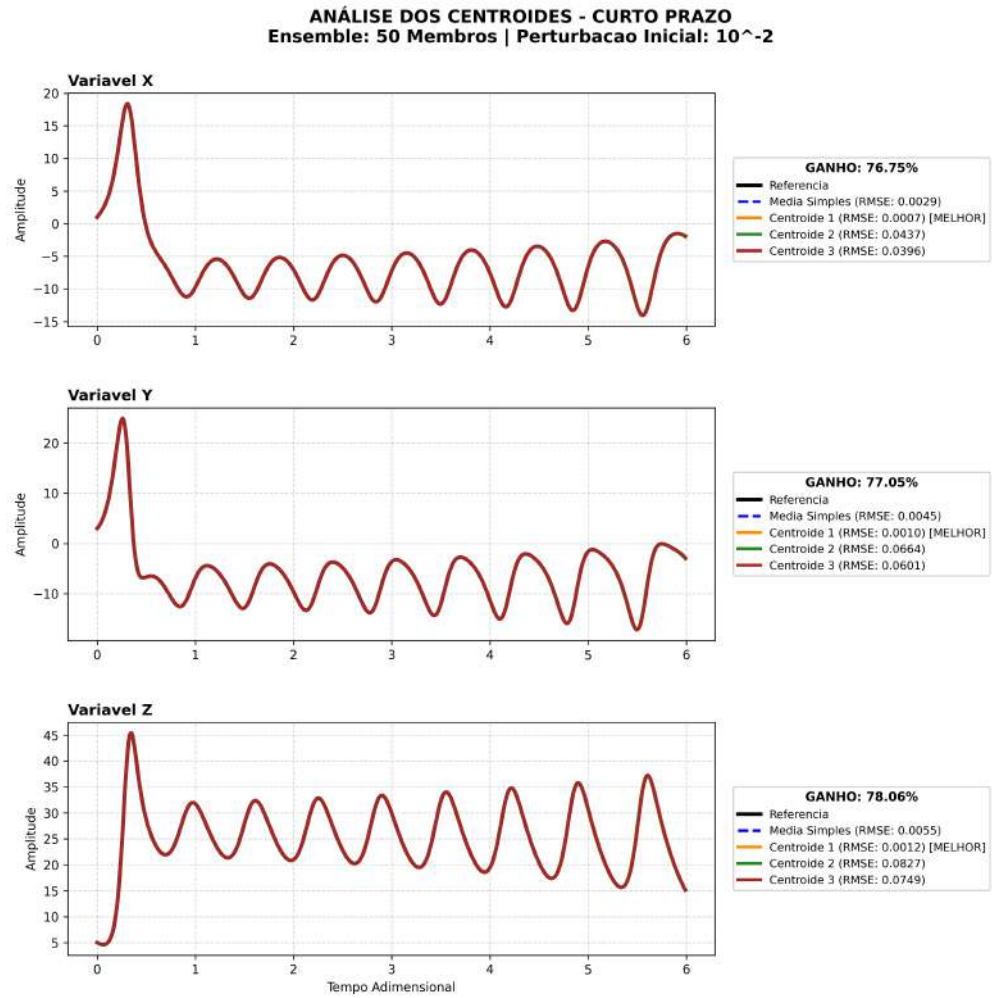


Figura 17 – Curto prazo ($0 \leq t < 6$), $N = 50$ e $\sigma_\eta = 10^{-2}$.

Fonte: Autoria própria.

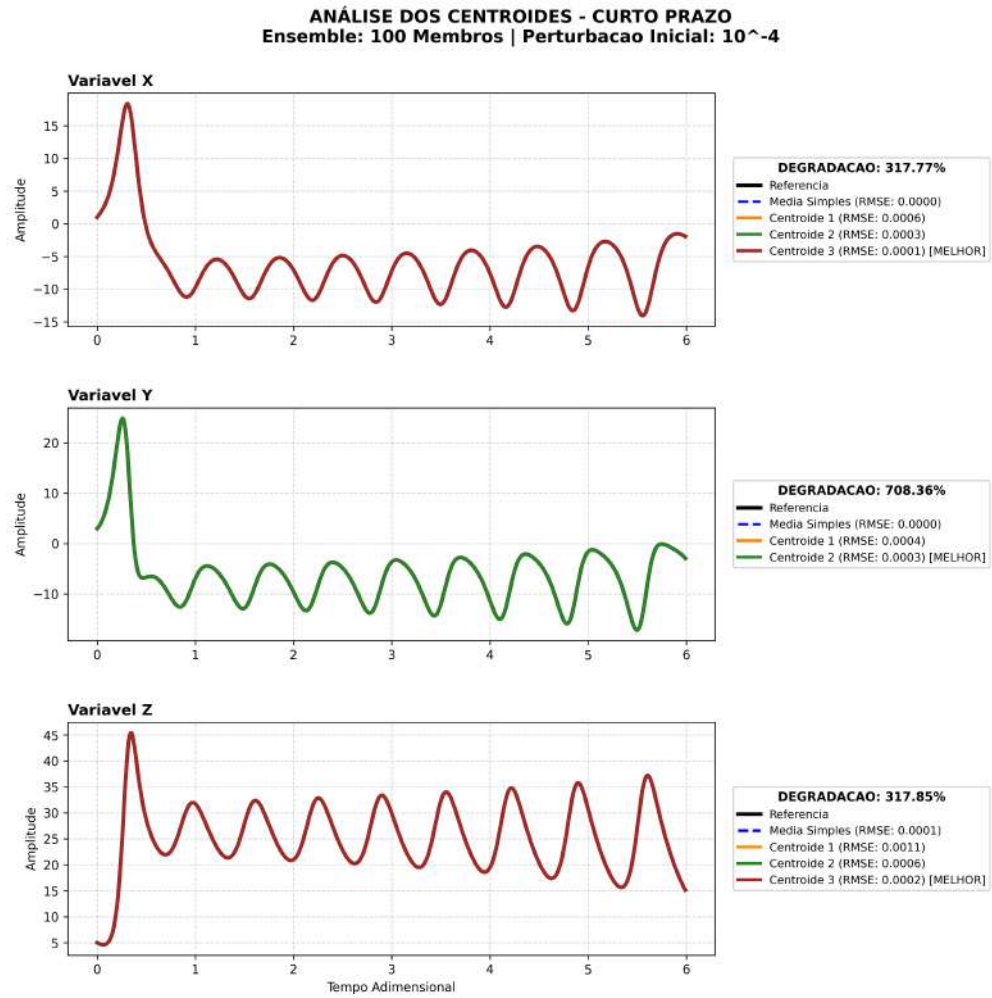


Figura 18 – Curto prazo ($0 \leq t < 6$), $N = 100$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

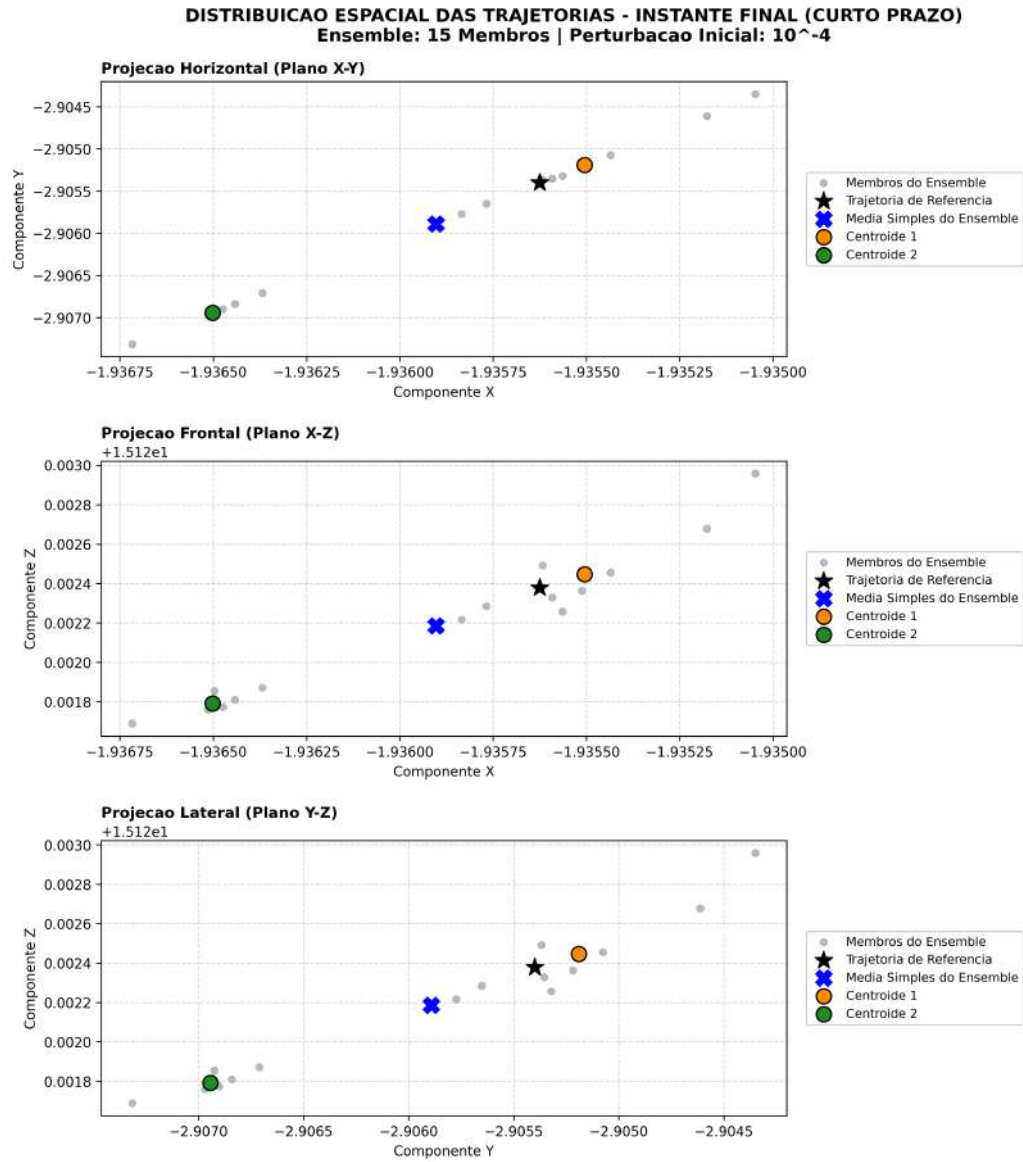


Figura 19 – Dispersão no curto prazo ($0 \leq t < 6$), $N = 15$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

Tabela complementar de resultados do horizonte temporal de curto prazo

Tabela 8 – Resultados do horizonte temporal de curto prazo ($0 \leq t < 6$).

N	σ_η	Var.	K	Silh.	RMSE M	RMSE C	Red. (%)
15	10^{-4}	X	2	0,6351	0,0001	0,0001	61,28
15	10^{-4}	Y	2	0,6324	0,0002	0,0001	60,92
15	10^{-4}	Z	2	0,6406	0,0003	0,0001	61,24
15	10^{-3}	X	2	0,6351	0,0014	0,0005	62,00
15	10^{-3}	Y	2	0,6323	0,0021	0,0008	61,65
15	10^{-3}	Z	2	0,6405	0,0026	0,0010	61,96
15	10^{-2}	X	2	0,6332	0,0136	0,0052	61,91
15	10^{-2}	Y	2	0,6305	0,0208	0,0080	61,57
15	10^{-2}	Z	2	0,6385	0,0258	0,0099	61,80
50	10^{-4}	X	3	0,5079	0,0000	0,0000	76,00
50	10^{-4}	Y	3	0,5034	0,0000	0,0000	76,17
50	10^{-4}	Z	3	0,5123	0,0001	0,0000	77,51
50	10^{-3}	X	3	0,5079	0,0003	0,0001	76,71
50	10^{-3}	Y	3	0,5034	0,0005	0,0001	76,91
50	10^{-3}	Z	3	0,5123	0,0006	0,0001	78,14
50	10^{-2}	X	3	0,5074	0,0029	0,0007	76,75
50	10^{-2}	Y	3	0,5030	0,0045	0,0010	77,05
50	10^{-2}	Z	3	0,5118	0,0055	0,0012	78,06
100	10^{-4}	X	3	0,4715	0,0000	0,0001	-317,77
100	10^{-4}	Y	2	0,4649	0,0000	0,0003	-708,36
100	10^{-4}	Z	3	0,4741	0,0001	0,0002	-317,85
100	10^{-3}	X	3	0,4713	0,0003	0,0011	-327,34
100	10^{-3}	Y	2	0,4649	0,0004	0,0034	-737,52
100	10^{-3}	Z	3	0,4739	0,0005	0,0022	-327,41
100	10^{-2}	X	3	0,4713	0,0028	0,0114	-309,03
100	10^{-2}	Y	3	0,4661	0,0043	0,0174	-307,88
100	10^{-2}	Z	3	0,4740	0,0053	0,0216	-307,10

Fonte: Autoria própria.

APÊNDICE B – Análise gráfica e tabelas complementares do horizonte temporal de médio prazo

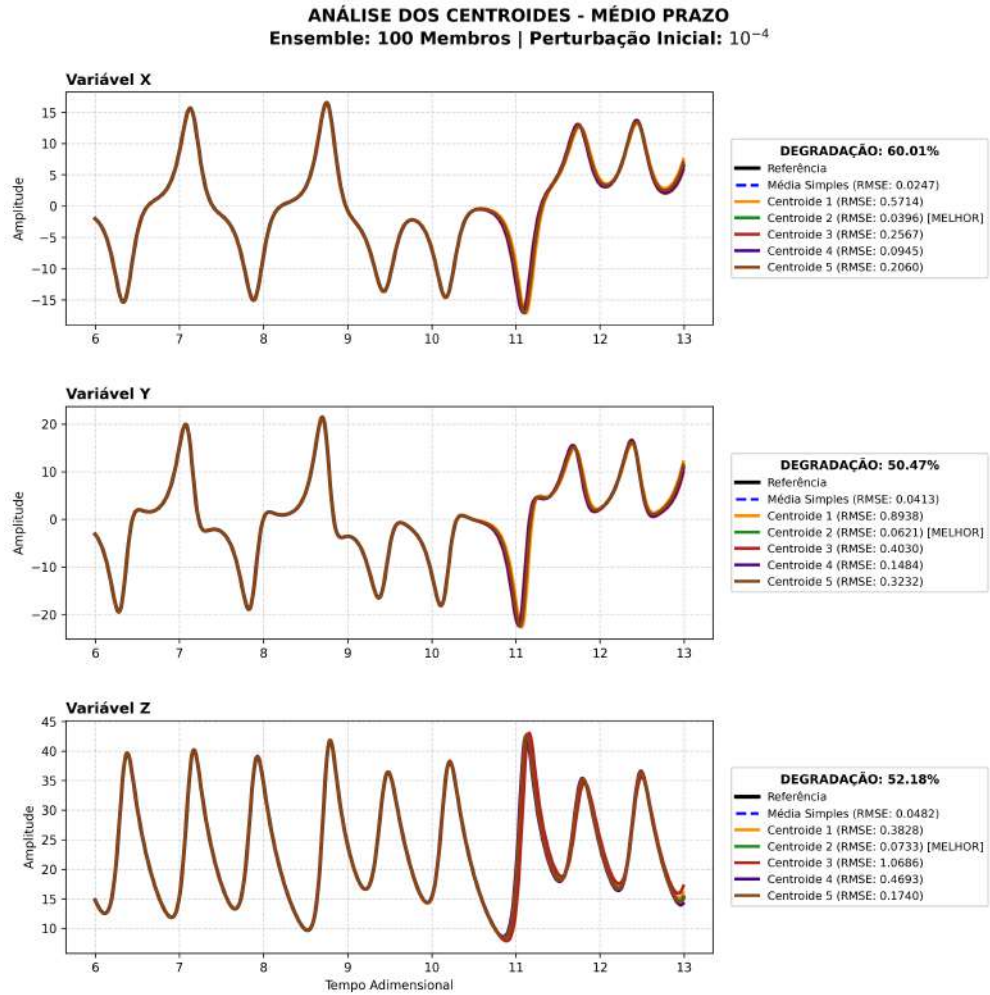


Figura 20 – Médio prazo ($6 \leq t < 13$), $N = 100$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

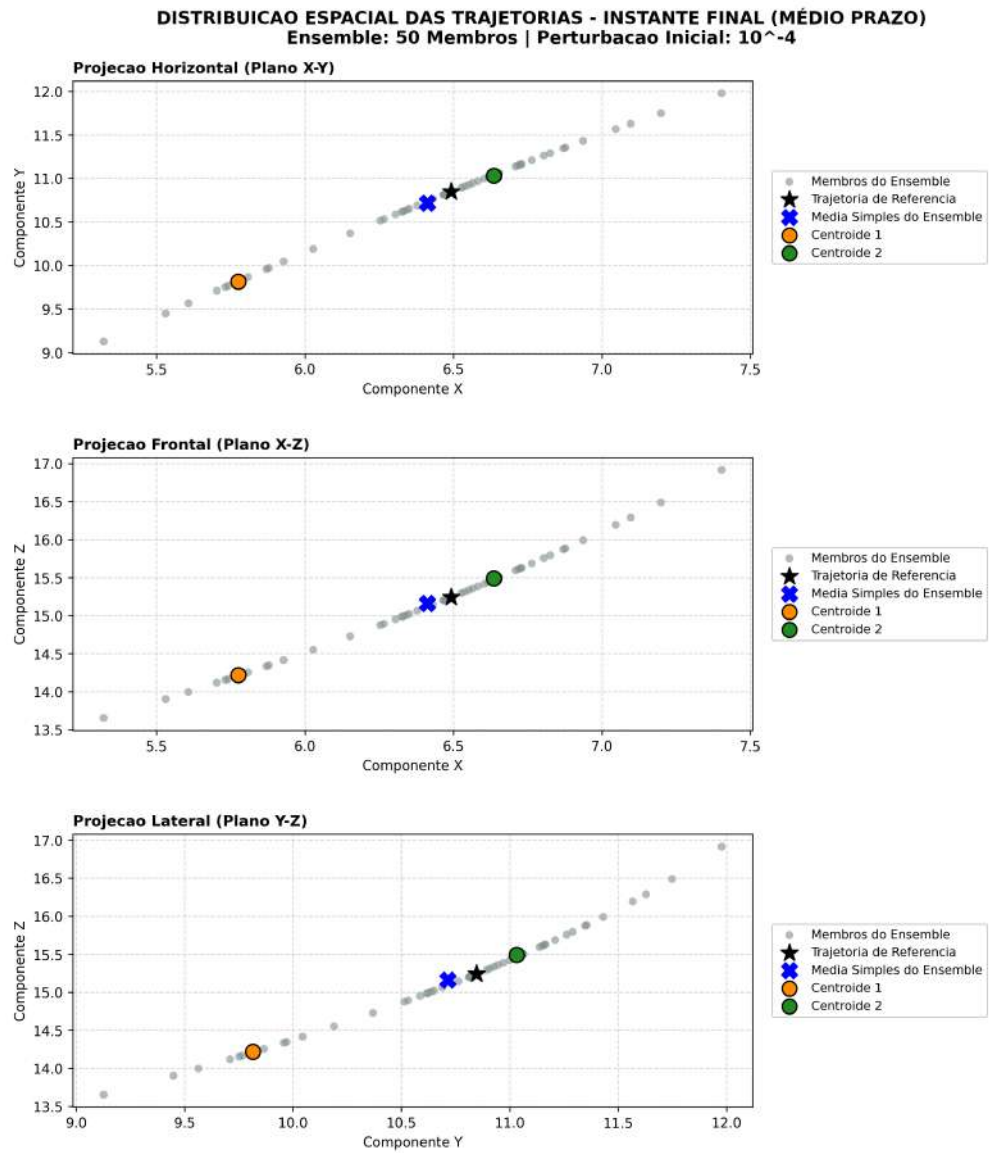


Figura 21 – Médio prazo ($6 \leq t < 13$), $N = 50$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

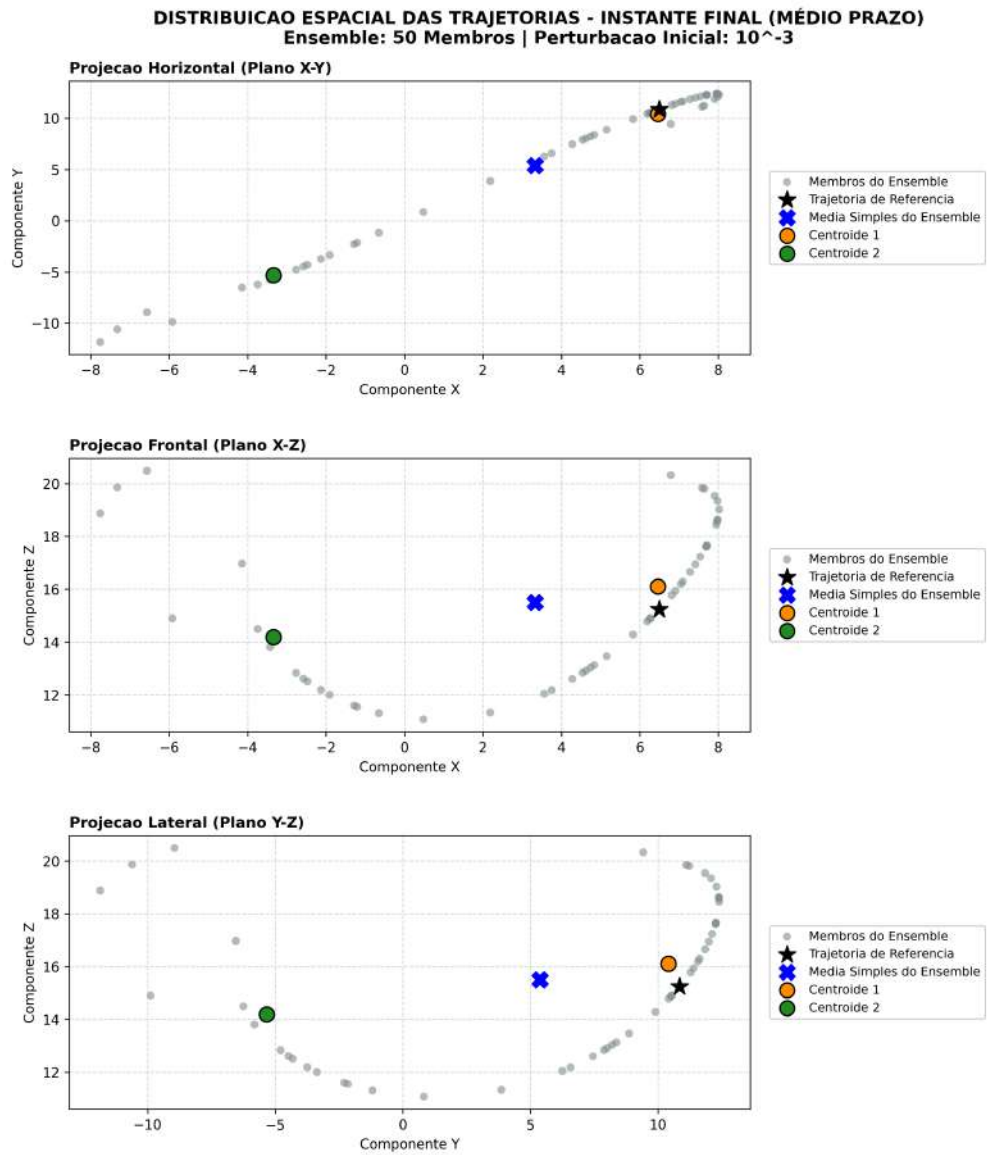


Figura 22 – Médio prazo ($6 \leq t < 13$), $N = 50$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

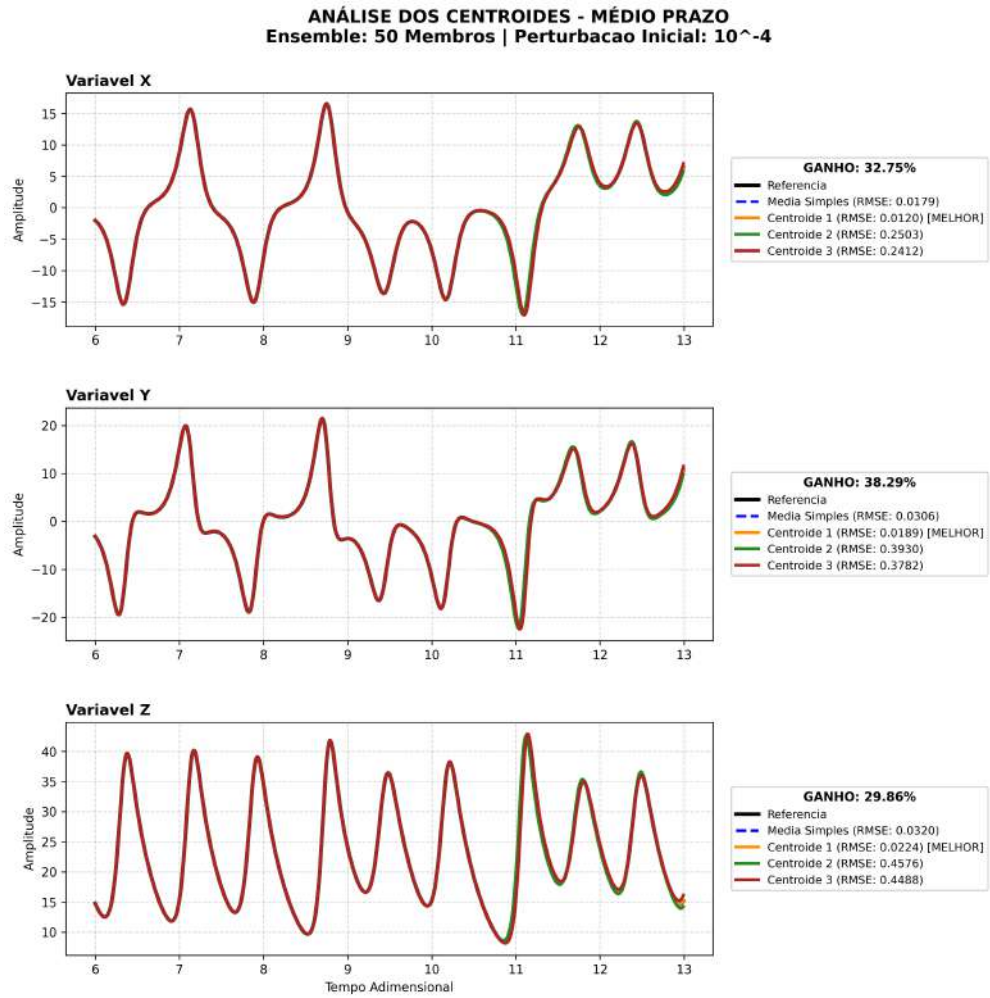


Figura 23 – Médio prazo ($6 \leq t < 13$), $N = 100$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

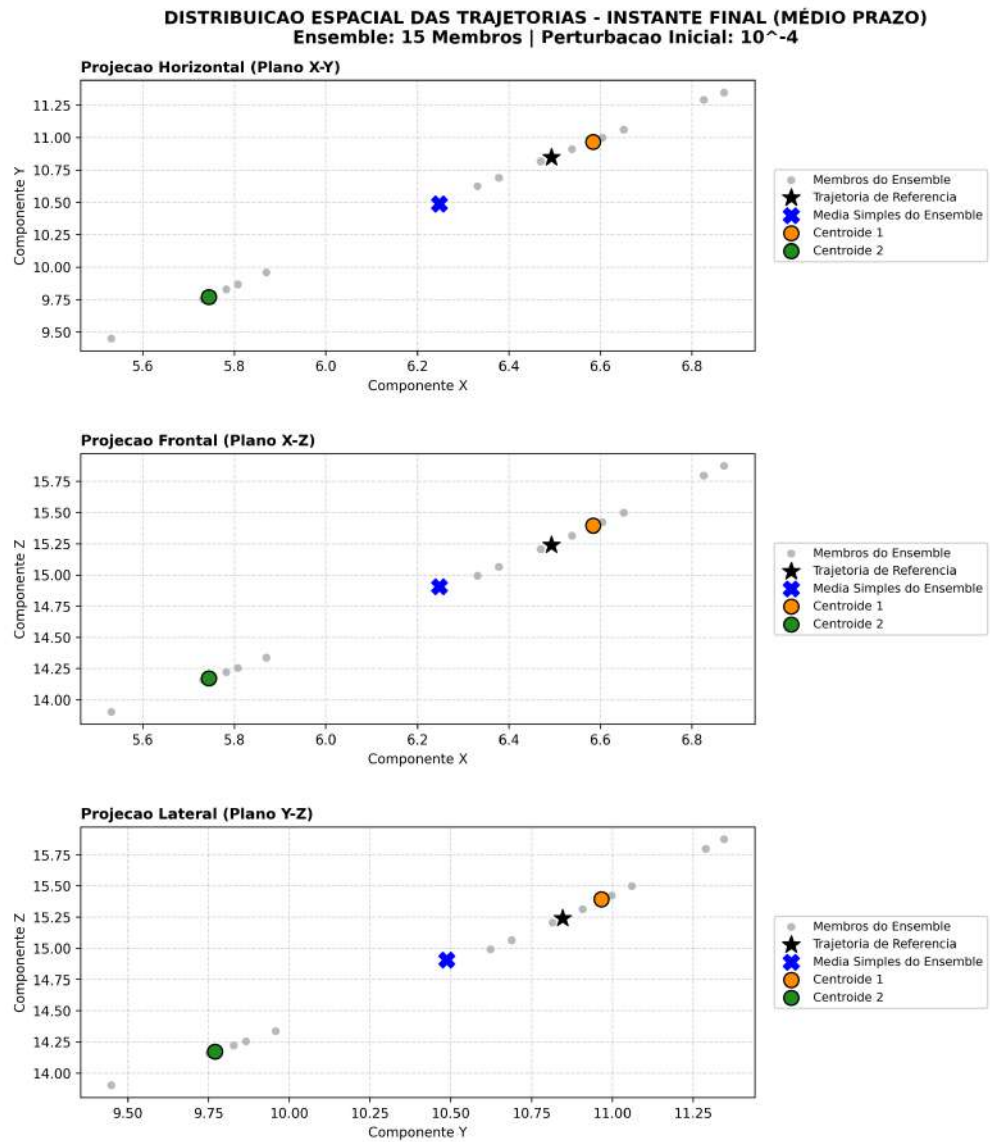


Figura 24 – Dispersão no médio prazo ($6 \leq t < 13$), $N = 15$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

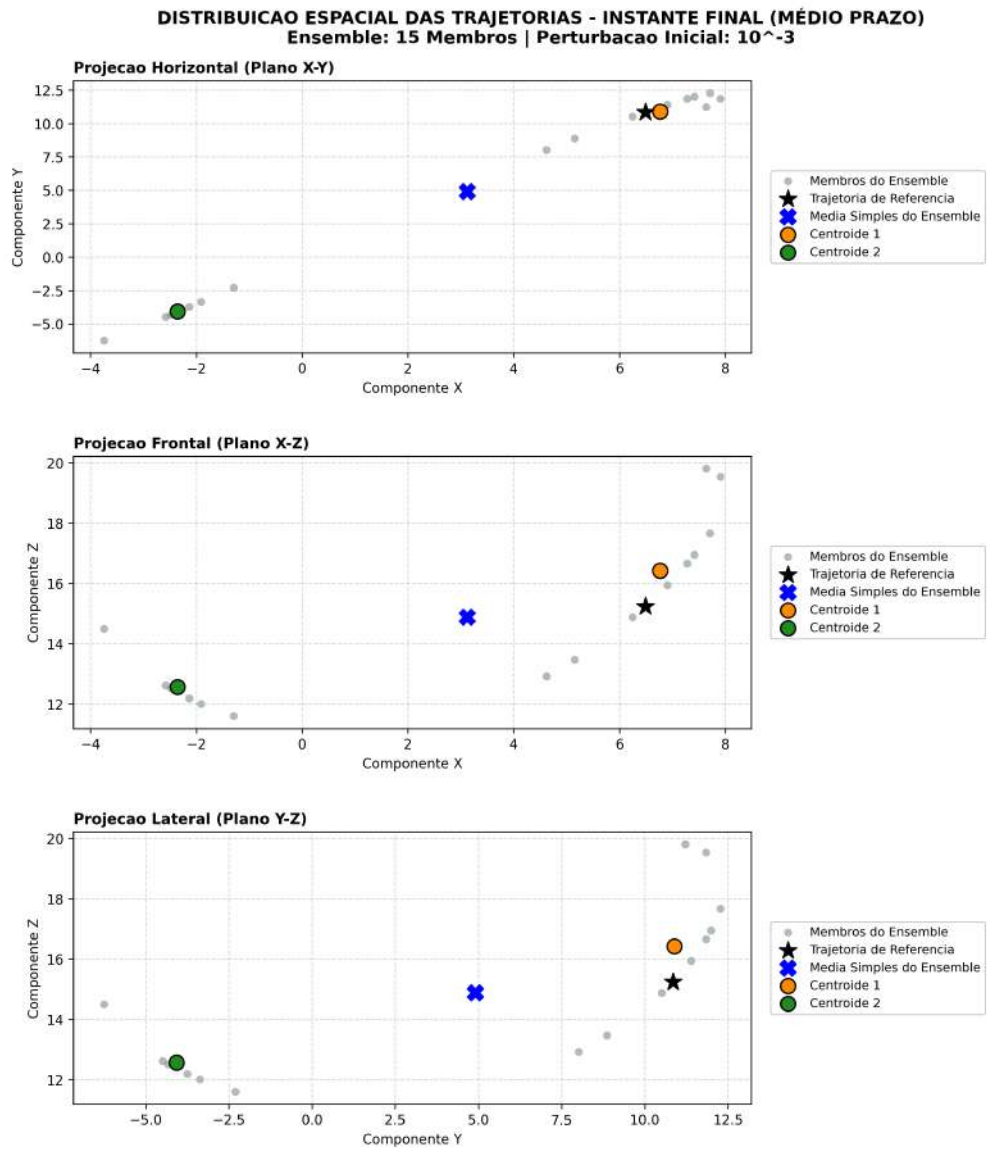


Figura 25 – Dispersão no médio prazo ($6 \leq t < 13$), $N = 15$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

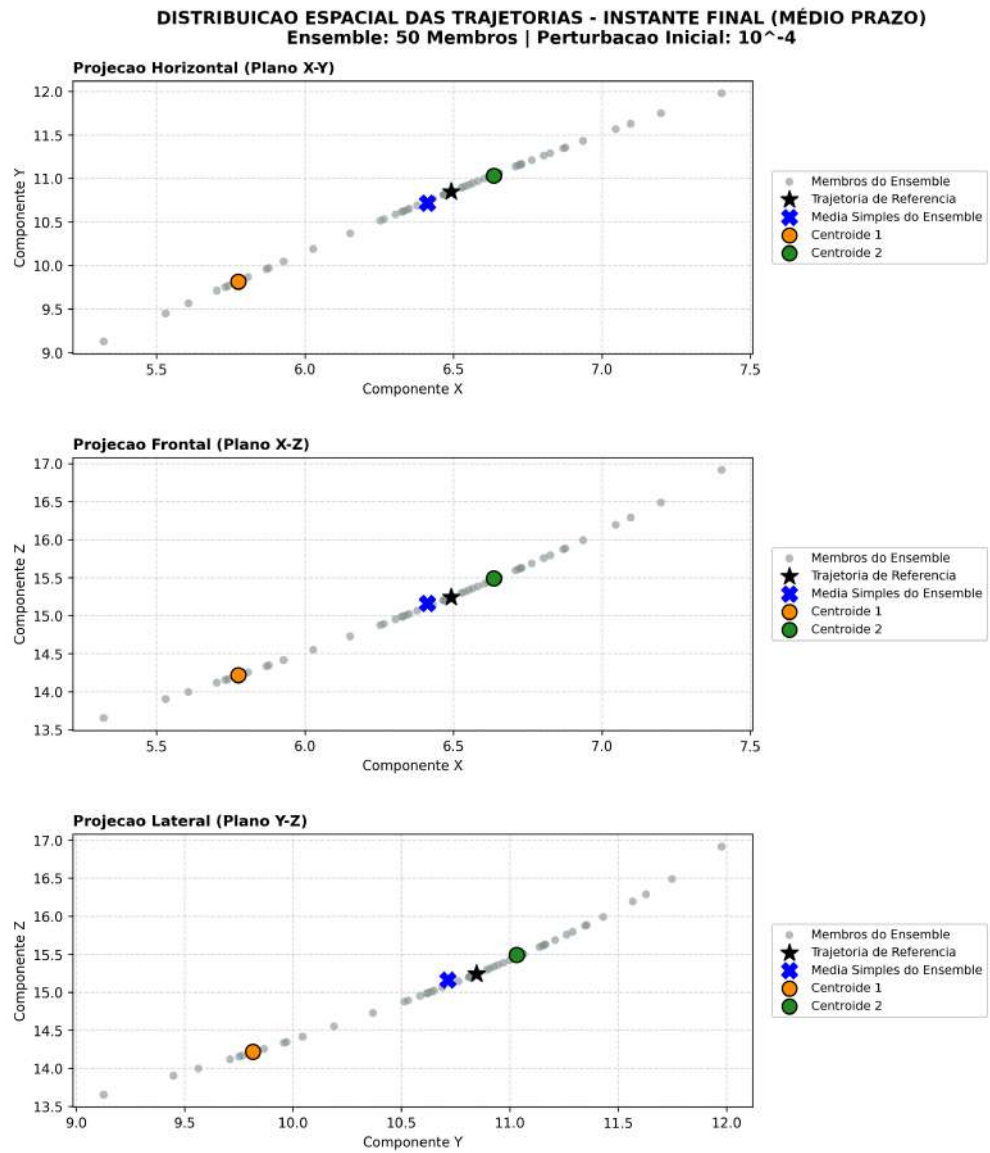


Figura 26 – Dispersão no médio prazo ($6 \leq t < 13$), $N = 50$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

Tabela complementar de resultados do horizonte temporal de médio prazo

Tabela 9 – Resultados do horizonte temporal de médio prazo ($6 \leq t < 13$).

N	σ_η	Var.	K	Silh.	RMSE M	RMSE C	Red. (%)
15	10^{-4}	X	2	0,7552	0,0808	0,0402	50,20
15	10^{-4}	Y	2	0,7551	0,1271	0,0631	50,32
15	10^{-4}	Z	2	0,7537	0,1471	0,0747	49,22
15	10^{-3}	X	3	0,7340	0,7123	0,1191	83,28
15	10^{-3}	Y	3	0,7305	1,2285	0,2005	83,68
15	10^{-3}	Z	3	0,7185	1,2612	0,2327	81,55
15	10^{-2}	X	5	0,4411	5,3145	2,0795	60,87
15	10^{-2}	Y	4	0,3983	6,2337	2,9549	52,60
15	10^{-2}	Z	2	0,4668	4,1193	2,7764	32,60
50	10^{-4}	X	3	0,5780	0,0179	0,0120	32,75
50	10^{-4}	Y	3	0,5778	0,0306	0,0189	38,29
50	10^{-4}	Z	3	0,5781	0,0320	0,0224	29,86
50	10^{-3}	X	2	0,8097	0,9763	0,5373	44,97
50	10^{-3}	Y	2	0,7444	1,3519	0,9461	30,01
50	10^{-3}	Z	4	0,5399	1,0910	0,3832	64,88
50	10^{-2}	X	5	0,4072	4,8940	1,6222	66,85
50	10^{-2}	Y	5	0,3497	5,6798	2,2193	60,93
50	10^{-2}	Z	2	0,4054	3,5776	3,1491	11,98
100	10^{-4}	X	5	0,5661	0,0247	0,0396	-60,01
100	10^{-4}	Y	5	0,5659	0,0413	0,0621	-50,47
100	10^{-4}	Z	5	0,5658	0,0482	0,0733	-52,18
100	10^{-3}	X	2	0,8024	1,3259	0,5117	61,41
100	10^{-3}	Y	2	0,7390	1,6750	0,9045	46,00
100	10^{-3}	Z	2	0,5638	1,2097	1,0765	11,01
100	10^{-2}	X	5	0,3933	4,6820	0,9125	80,51
100	10^{-2}	Y	5	0,3410	5,4898	2,3850	56,55
100	10^{-2}	Z	5	0,3706	3,6163	1,3654	62,24

Fonte: Autoria própria.

APÊNDICE C – Análise gráfica e tabelas complementares do horizonte temporal de longo prazo

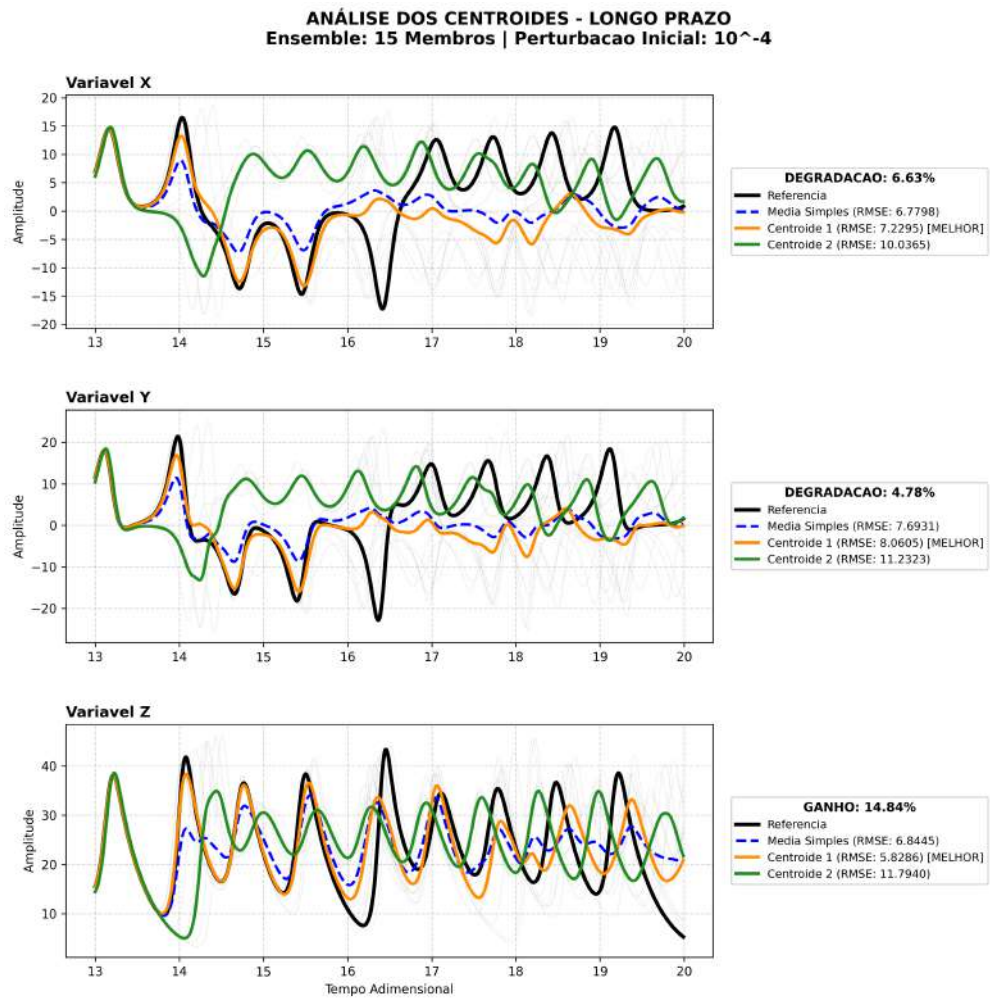


Figura 27 – Longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

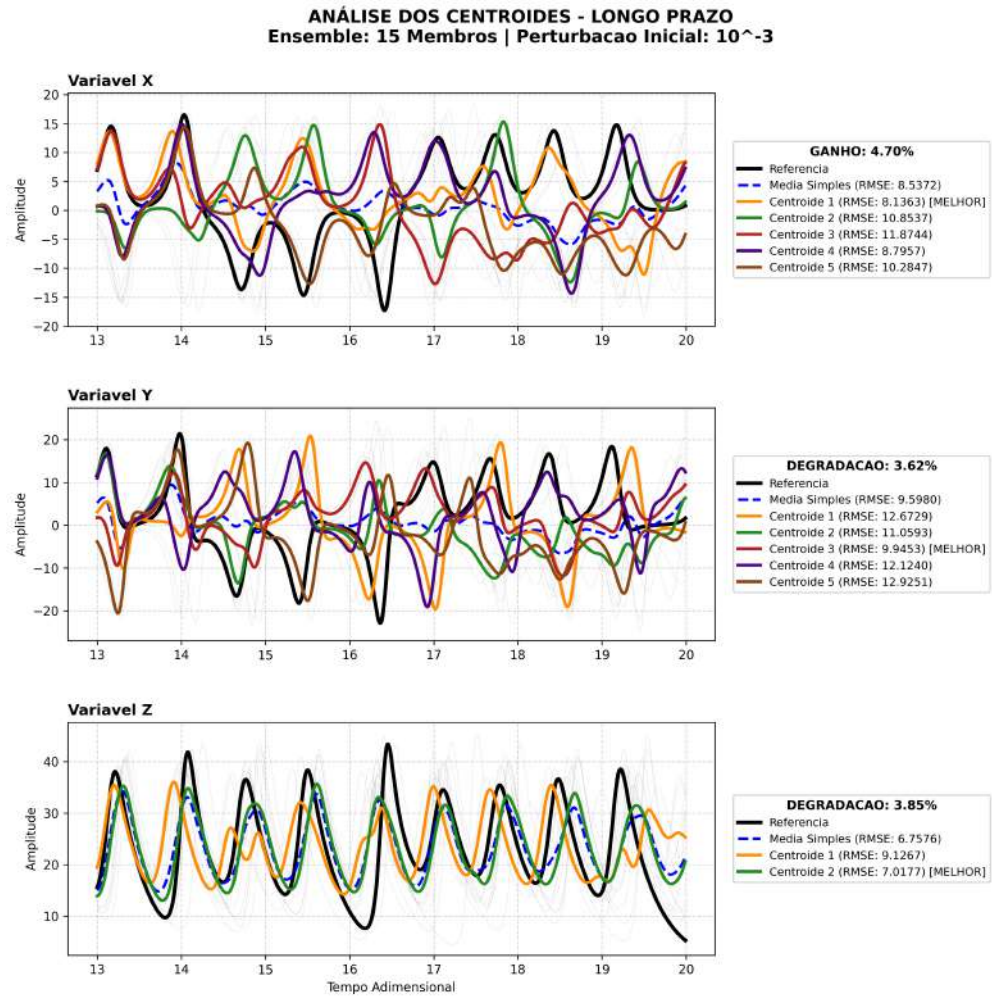


Figura 28 – Longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

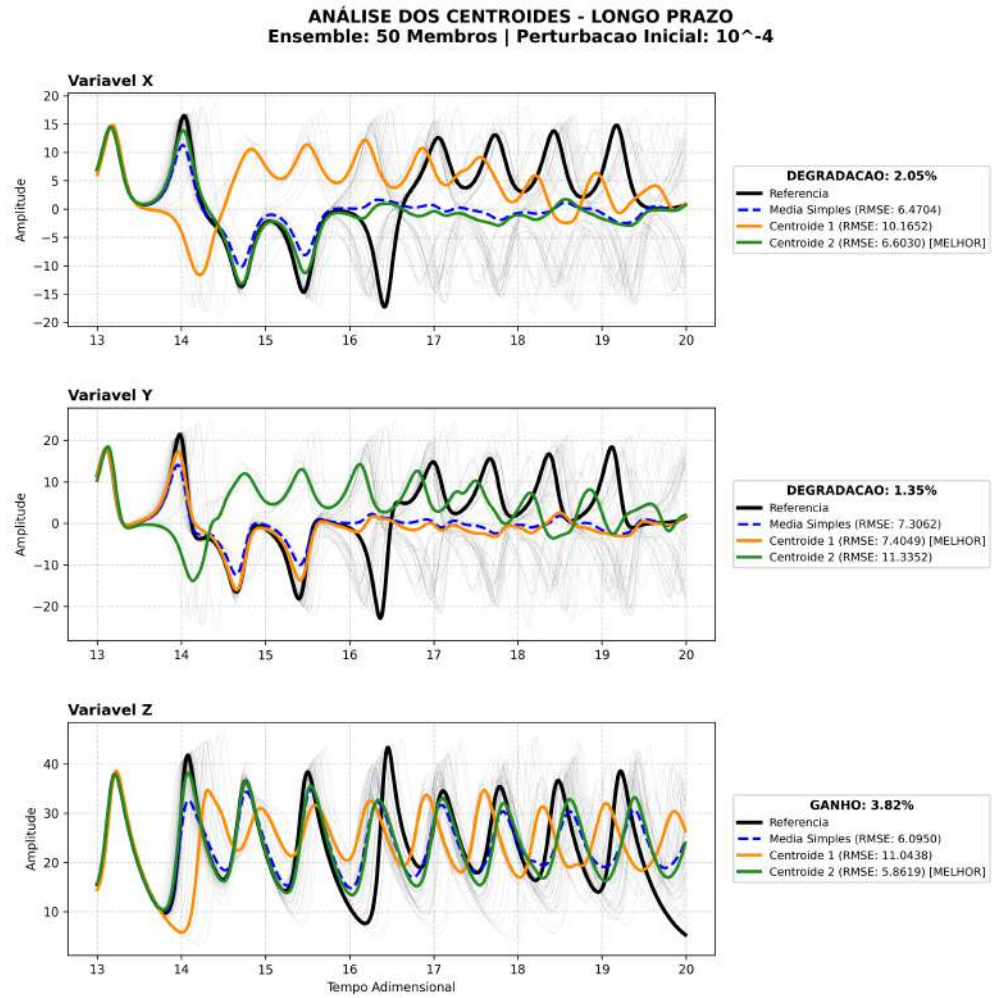


Figura 29 – Longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

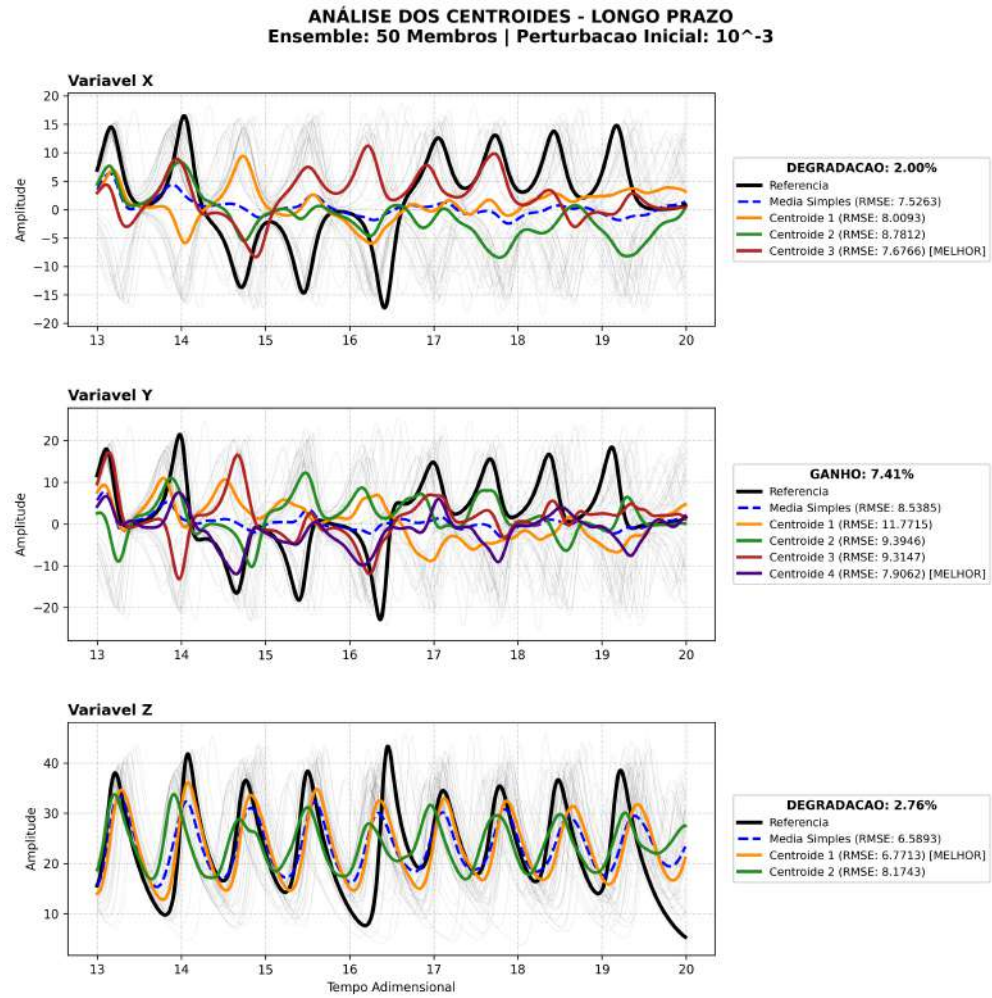


Figura 30 – Longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

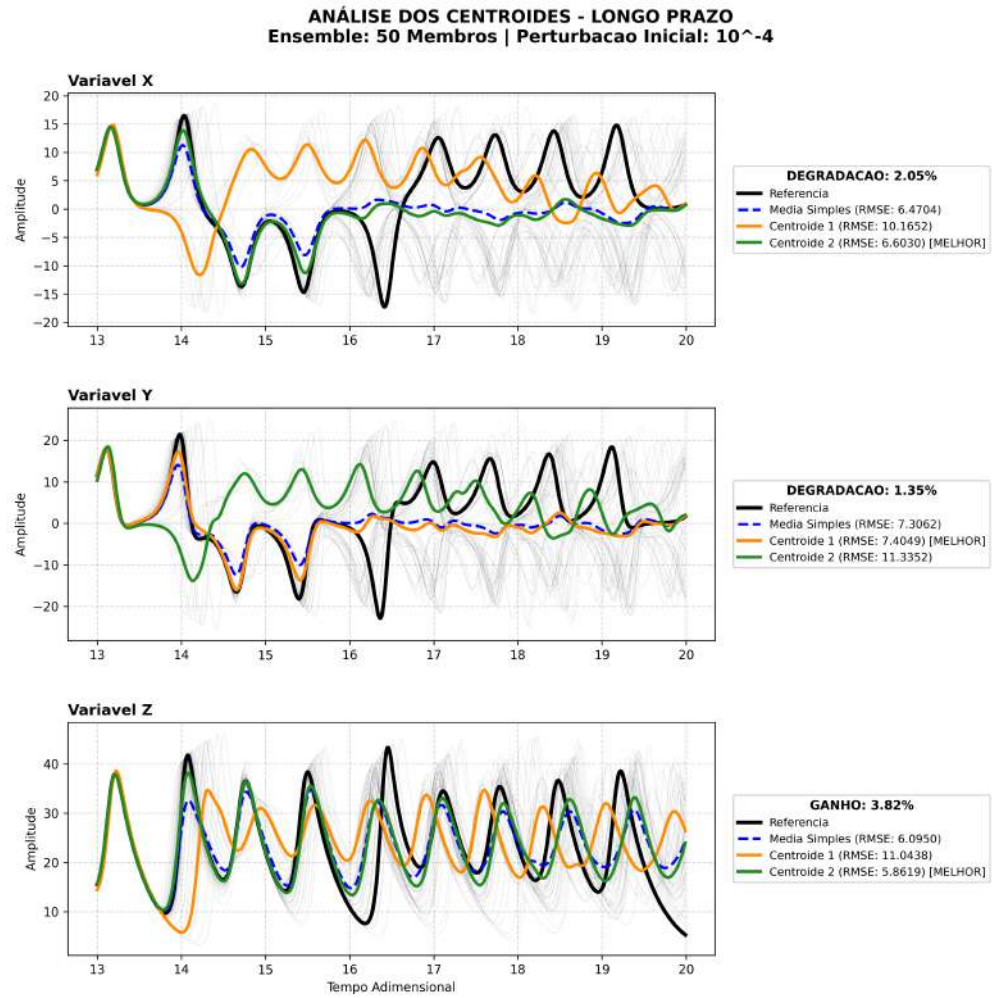


Figura 31 – Longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

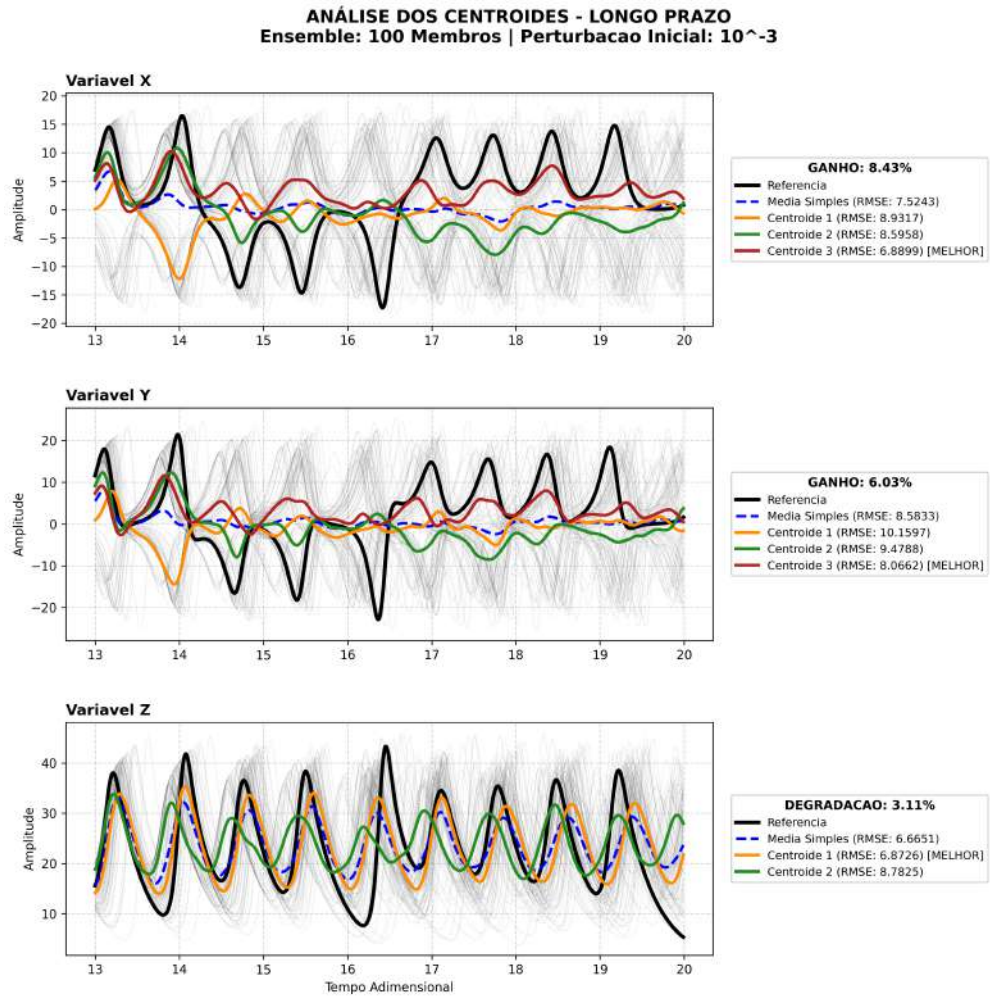


Figura 32 – Longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

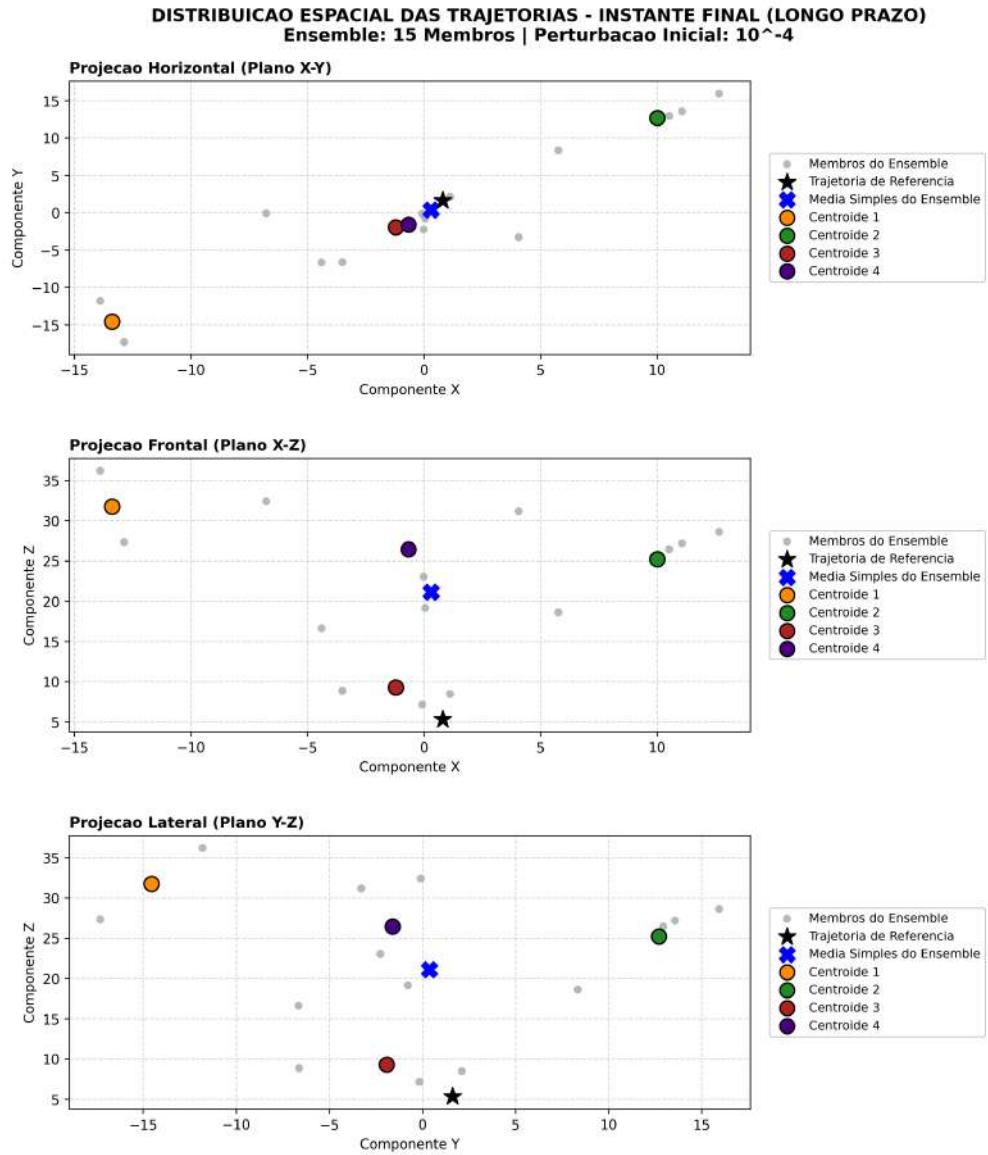


Figura 33 – Dispersão no longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

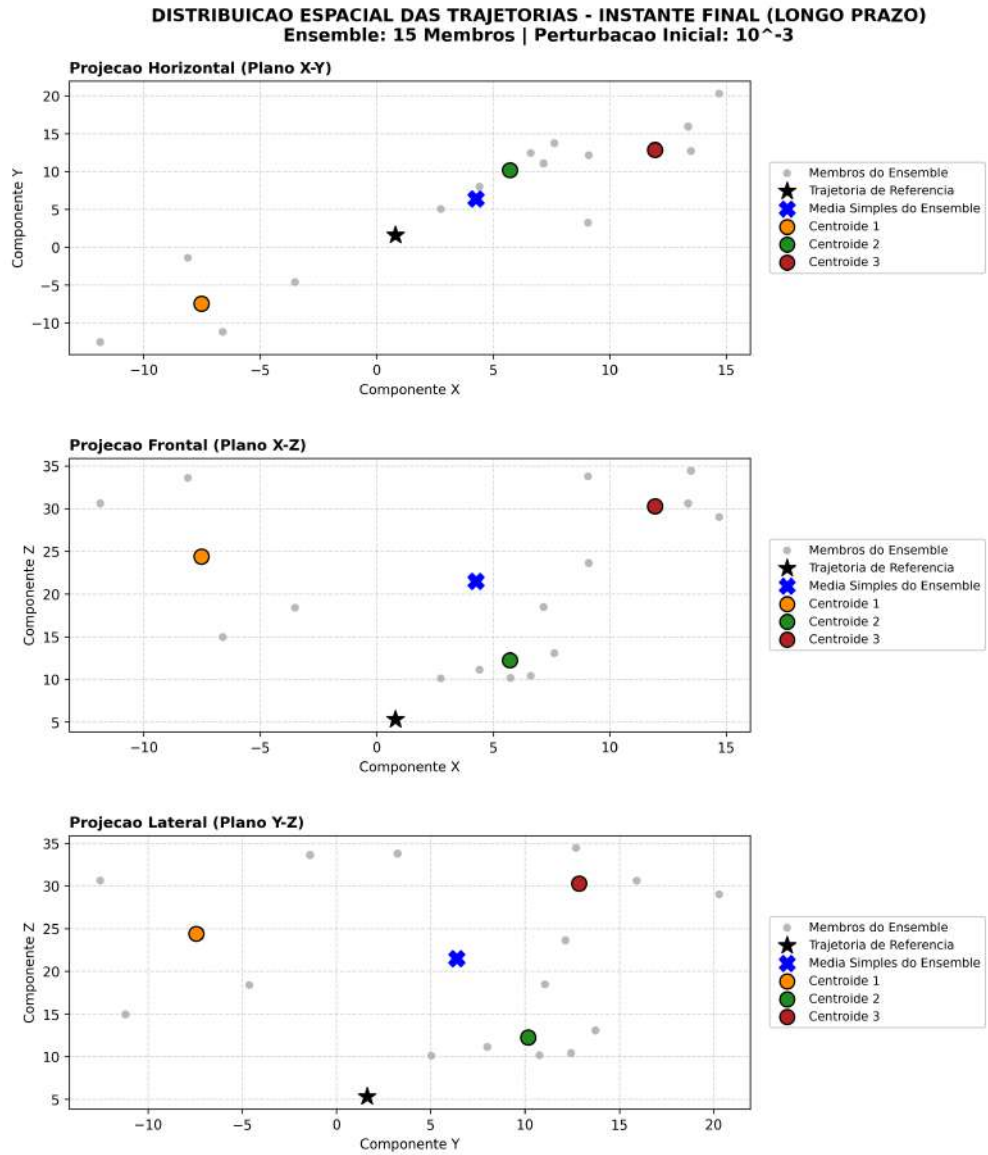


Figura 34 – Dispersão no longo prazo ($13 \leq t < 20$), $N = 15$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

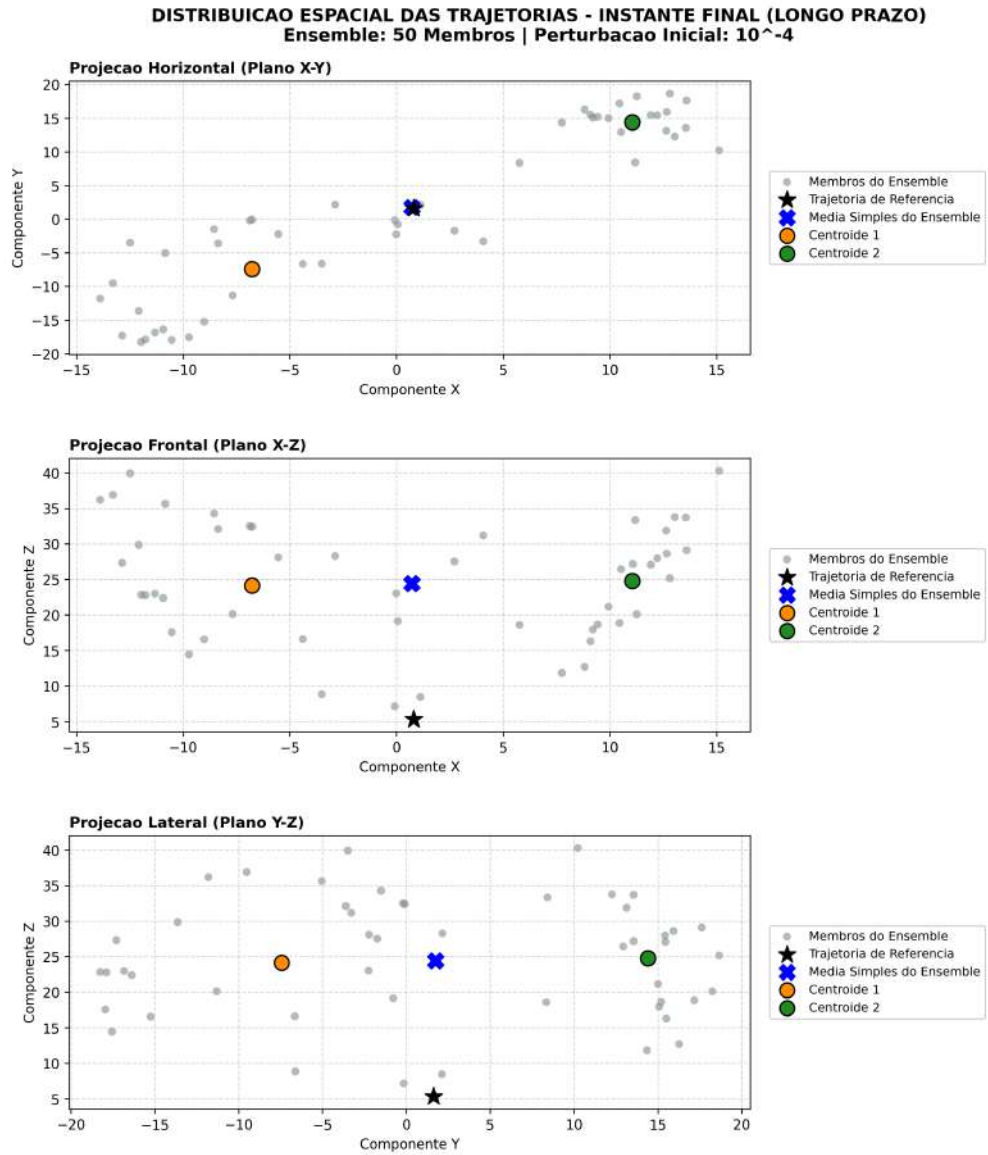


Figura 35 – Dispersão no longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

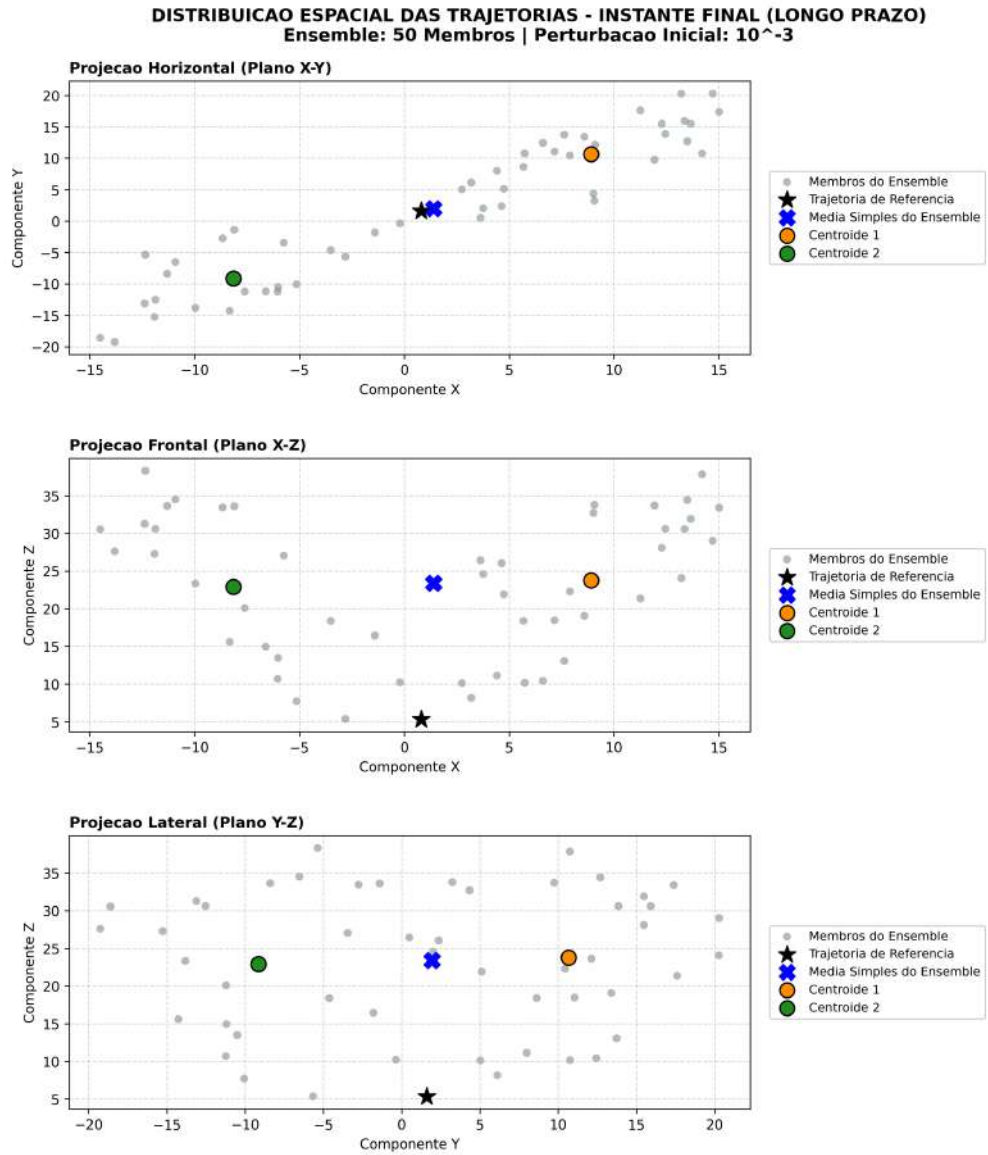


Figura 36 – Dispersão no longo prazo ($13 \leq t < 20$), $N = 50$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

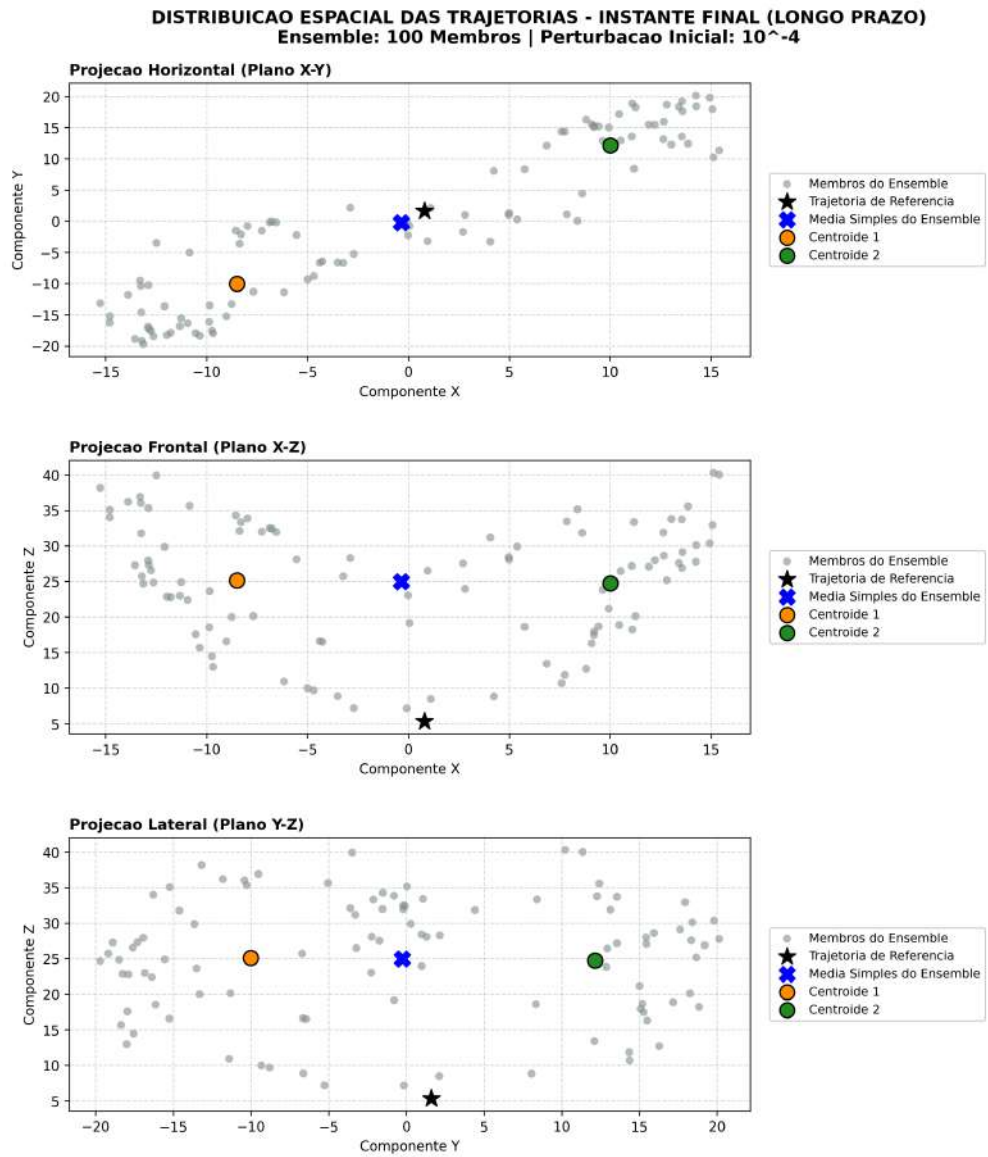


Figura 37 – Dispersão no longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-4}$.

Fonte: Autoria própria.

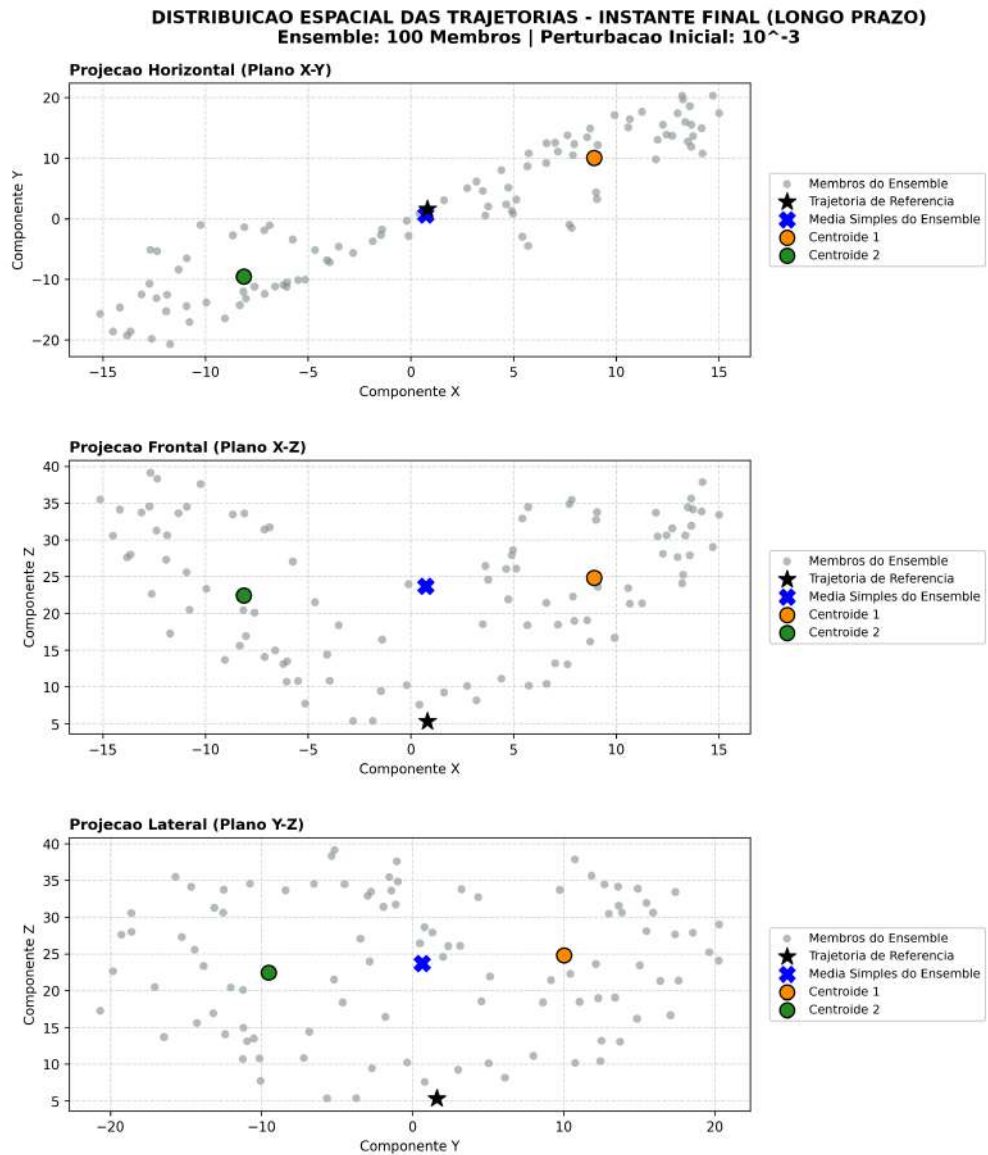


Figura 38 – Dispersão no longo prazo ($13 \leq t < 20$), $N = 100$ e $\sigma_\eta = 10^{-3}$.

Fonte: Autoria própria.

Tabela complementar de resultados do horizonte temporal de longo prazo

Tabela 10 – Resultados do horizonte temporal de longo prazo ($13 \leq t < 20$).

N	σ_η	Var.	K	Silh.	RMSE M	RMSE C	Red. (%)
15	10^{-4}	X	2	0,3386	6,7798	7,2295	-6,63
15	10^{-4}	Y	2	0,2991	7,6931	8,0605	-4,78
15	10^{-4}	Z	2	0,3954	6,8445	5,8286	14,84
15	10^{-3}	X	5	0,1349	8,5372	8,1363	4,70
15	10^{-3}	Y	5	0,1294	9,5980	9,9453	-3,62
15	10^{-3}	Z	2	0,2098	6,7576	7,0177	-3,85
15	10^{-2}	X	5	0,1255	8,1319	9,2079	-13,23
15	10^{-2}	Y	5	0,1230	9,2252	9,7395	-5,58
15	10^{-2}	Z	2	0,2430	7,4478	7,5799	-1,77
50	10^{-4}	X	2	0,2863	6,4704	6,6030	-2,05
50	10^{-4}	Y	2	0,2515	7,3062	7,4049	-1,35
50	10^{-4}	Z	2	0,3658	6,0950	5,8619	3,82
50	10^{-3}	X	3	0,0977	7,5263	7,6766	-2,00
50	10^{-3}	Y	4	0,0905	8,5385	7,9062	7,41
50	10^{-3}	Z	2	0,1802	6,5893	6,7713	-2,76
50	10^{-2}	X	3	0,0892	7,6248	6,3849	16,26
50	10^{-2}	Y	4	0,0719	8,6514	7,6703	11,34
50	10^{-2}	Z	2	0,2561	7,2919	7,4412	-2,05
100	10^{-4}	X	5	0,1744	6,1514	4,6738	24,02
100	10^{-4}	Y	4	0,1543	7,0073	4,6961	32,98
100	10^{-4}	Z	2	0,3286	5,9474	5,7880	2,68
100	10^{-3}	X	3	0,0929	7,5243	6,8899	8,43
100	10^{-3}	Y	3	0,0811	8,5833	8,0662	6,03
100	10^{-3}	Z	2	0,2308	6,6651	6,8726	-3,11
100	10^{-2}	X	2	0,0960	7,8771	7,4880	4,94
100	10^{-2}	Y	2	0,0835	8,8955	8,0317	9,71
100	10^{-2}	Z	2	0,2129	7,0979	7,3982	-4,23

Fonte: Autoria própria.