PINNeAPPle

RELATÓRIO DE ESTUDO · EXEMPLO PÚBLICO

El Niño
trinta configurações, um protocolo só

Modelos físicos, de dados, physics-informed e híbridos avaliados com a mesma métrica, as mesmas janelas e o mesmo teste de significância.

0,95

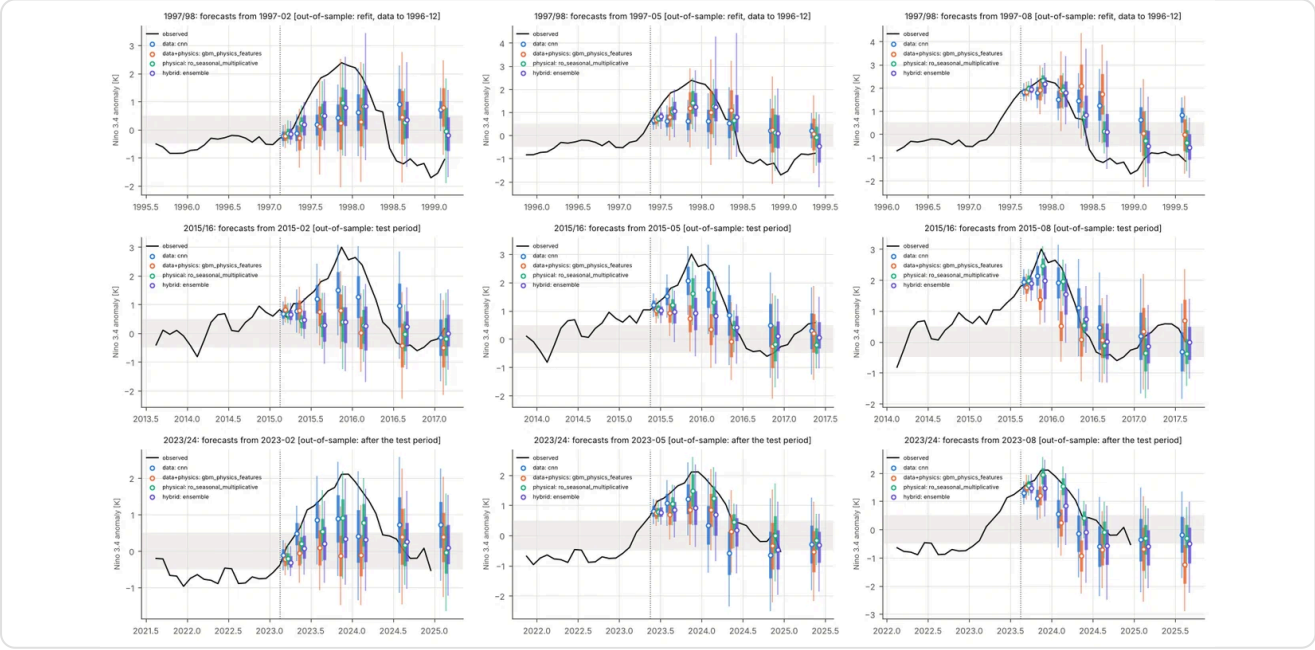
de ACC em h = 1 no melhor modelo

3

configurações com ganho significativo sobre o melhor baseline

−0,69

de ACC no PINN com estimador de estado em h = 3



Documento GEO-01 · versão 1.0 · 2026
Dados abertos: OISST v2.1, WOD, ERA5 (cache com SHA-256 de cada arquivo)
Protocolo único: ACC, skill contra persistência e contra o melhor baseline, Holm
PINNeAPPle · LAIS & YAN LTDA · pinneapple-labs.github.io

Prever o El Niño é o teste clássico de previsão de sistemas físicos: um ciclo irregular, um oceano acoplado à atmosfera e uma barreira de primavera em que a habilidade de todos os modelos cai. É também o lugar onde “meu modelo é melhor” costuma ser dito sem protocolo.

Este estudo coloca trinta configurações no mesmo banco: baselines estatísticos, modelos de dados, modelos físicos, physics-informed e híbridos. Todos veem as mesmas janelas, são avaliados nos mesmos horizontes e passam pelo mesmo teste de significância.

HORIZONTE	O QUE MUDA
h = 1 a 3 meses	persistência e climatologia ainda são fortes; a diferença entre modelos é pequena
h = 6 meses	atravessa a barreira da primavera: é onde o protocolo separa
h = 12 a 24 meses	quase todos perdem habilidade; quem ainda ganha, ganha pouco

> **A regra do banco.** Nenhum modelo escolhe a métrica, a janela ou o baseline. Quem reporta “ACC 0,95” sem dizer contra o que, em que horizonte e com qual significância não reportou nada.

ITEM	DEFINIÇÃO
Dados	OISST v2.1, WOD e ERA5; cada arquivo entra com o SHA-256 registrado
Métrica principal	ACC (correlação de anomalia) por horizonte
Skills	contra persistência e contra o melhor baseline em cada horizonte
Significância	teste de Holm sobre as diferenças, não sobre a média
Incerteza	cobertura do intervalo de 90% e desvio de PIT

As quatro famílias entram com representantes escolhidos na validação, não no teste — o conjunto de teste é medido uma vez, depois da escolha.

CONFIGURAÇÃO	FAMÍLIA	ACC H=1	ACC H=3	ACC H=6	ACC H=12	SIGNIFICATIVO
persistência	baseline	0,93	0,73	0,36	0,01	—
arima	baseline	0,93	0,76	0,42	0,02	não
cnn	dados	0,95	0,87	0,70	0,26	h = 1
cnn pré-treinada	dados+física	0,95	0,85	0,66	0,30	h = 1, 2, 3
ro sazonal multiplicativo	físico	0,95	0,84	0,63	0,30	não
ensemble	híbrido	0,95	0,84	0,58	0,20	h = 1, 2, 3
pinn_ro (estimador de estado)	dados+física	0,94	−0,69	−0,35	−0,01	não
climatologia	baseline	0,00	0,00	0,00	0,00	—

Em h = 1 quase tudo empata perto de 0,95: nesse horizonte a persistência já é boa, e a diferença entre modelos é ruído. Em h = 6 a separação aparece — e só **três** configurações têm ganho significativo sobre o melhor baseline.

> **O resultado que ninguém gosta de publicar.** O PINN com estimador de estado (`pinn_ro`) cai para ACC **−0,69** em h = 3: pior que a média do próprio período. Entra no relatório porque entrou no banco.

FAMÍLIA	REPRESENTANTE	SKILL DE CRPS VS CLIMATOLOGIA (H = 1, 3, 6)	COBERTURA MÉDIA	DESVIO DE PIT
baseline	arima	0,64 · 0,32 · 0,06	0,94	0,20
dados	cnn	0,69 · 0,45 · 0,09	0,98	0,28
físico	ro sazonal	0,67 · 0,44 · 0,20	0,88	0,17
híbrido	ensemble	0,68 · 0,45 · 0,19	0,95	0,12
dados+física	gbm com features físicas	0,64 · 0,28 · −0,08	0,92	0,23

A cobertura de 0,98 do modelo de dados parece ótima até se olhar o PIT de 0,28: os intervalos são largos demais. O híbrido com PIT de 0,12 é o mais calibrado do conjunto — e não é o de maior ACC.

> **Duas perguntas, não uma.** “Qual acerta mais” e “qual sabe o quanto erra” não têm o mesmo vencedor. Quem escolhe modelo por ACC escolhe metade da resposta.

ITEM	O QUE O RELATÓRIO DIZ
Barreira da primavera	medida separadamente; é onde a habilidade cai para todos
Horizonte longo	acima de 12 meses, o ganho é pequeno e some em várias configurações
Dados	OISST, WOD e ERA5 com SHA-256; fora dessas fontes o protocolo não vale
Escolha de modelo	feita na validação; teste e extrapolação nunca entram na escolha
PINNs	duas variantes com estimador de estado falham de forma visível — declarado
Aplicação	previsão de índice, não de impacto regional (chuva, safra, energia)

O que este estudo não faz. Não prevê chuva nem safra: prevê o índice do El Niño. A passagem do índice para impacto regional é outro estudo, com outros dados e outros erros.

0,95
de ACC em h = 1

3
configurações significativas

0,20
de ACC em h = 6 na cnn

0,12
de desvio de PIT no ensemble

O que dá para decidir

Qual família usar para cada horizonte, com que incerteza e com que ganho real sobre o que já existe. Em h = 1 a resposta é “use o mais barato”; em h = 6, só três configurações se sustentam.

Proveniência

O QUE	ONDE
Tabela completa	`PINNeAPPle/examples/enso_forecasting/results/tables.md`
Estudo	`results/study.json · phase.json · spring_barrier.json`
Dados	OISST v2.1, WOD, ERA5 — com SHA-256 no cache
Artigo público	`pinneapple-labs.github.io/artigos/el-nino-28-modelos/`

Este é um relatório de exemplo, sobre dados abertos e um protocolo reproduzível. Um estudo para o seu sistema segue a mesma estrutura: baseline explícito, teste de significância e a página do que não foi previsto.