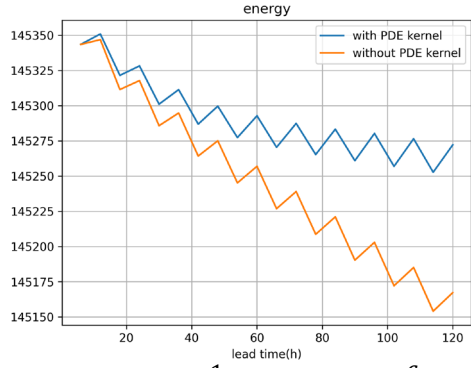


(a-c) bias =  $\overline{pred} - gt$  (the closer to 0 the better) [1]

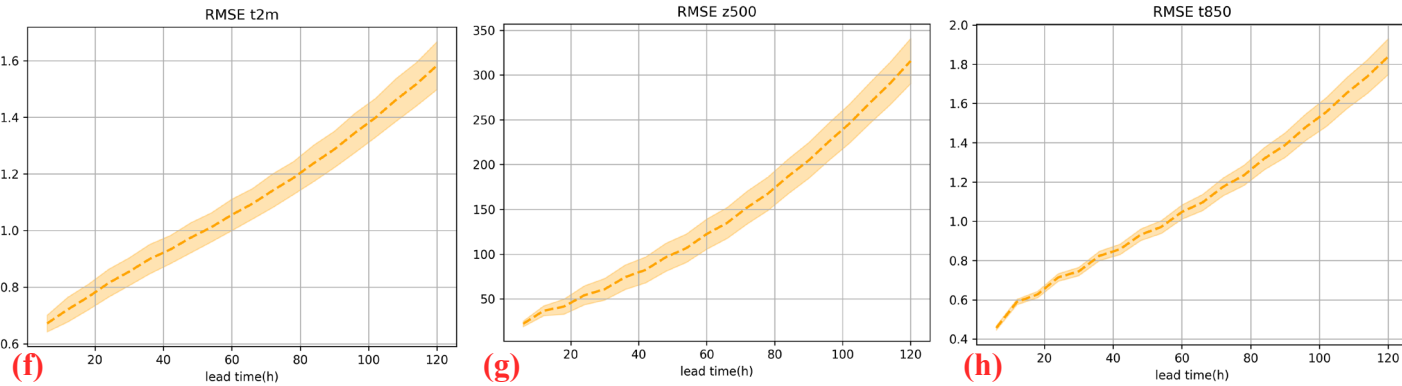


(d) energy =  $\frac{1}{2}(u^2 + v^2) + \frac{c_p}{2T_r}T^2$

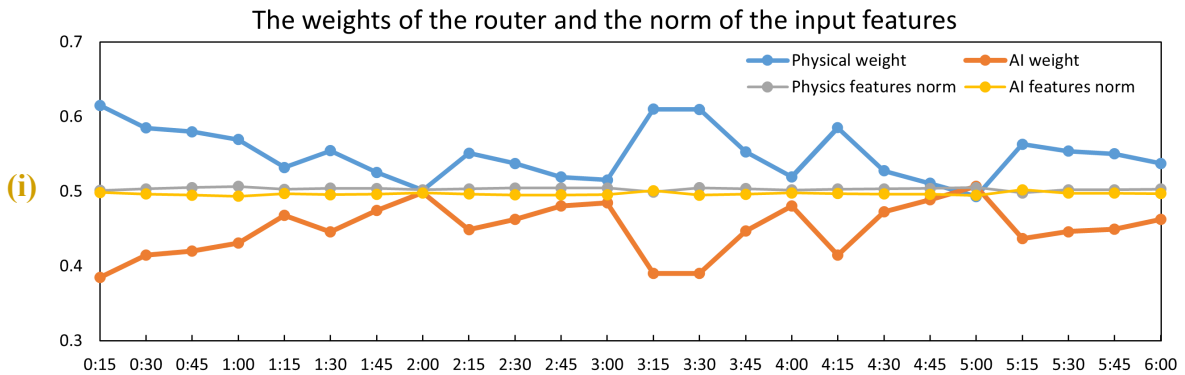
(the more consistent the better) [2]

Model	Cost time(s) per training step
FourCastNet	1.8
ClimODE	2.5
Keisler	3.1
WeatherGFT without PDE kernel	2.2
WeatherGFT (ours)	2.4

(e) Comparison of time consumption



(f-h) RMSE Error Bars



(i) The norms of the outputs from the two networks are similar and stable. This indicates: (I) The two networks produce outputs on the same scale. (II) The router is decoupled and dynamically selects the more crucial features from the two branches without affecting the scale of the two networks.