Bridging Traditional Econometrics and Artificial Intelligence in Financial Forecasting:

A Novel Approach to Stock Return Prediction

Georgiou Catherine[a*], Fassas Athanasios[b], Gkonis Vasileios[c] and Tsakalos, Ioannis[d]

[a]Post-Doc researcher, Department of Accounting and Finance, University of Thessaly, aikategeorgiou@uth.gr

[b]Professor, Department of Accounting and Finance, University of Thessaly

[c]cPhD, Department of Accounting and Finance, University of Thessaly

[d]Assistant Professor, Department of Accounting and Finance, University of Thessaly

**Corresponding author*

Introduction

Why use machine learning techniques for forecasting purposes?

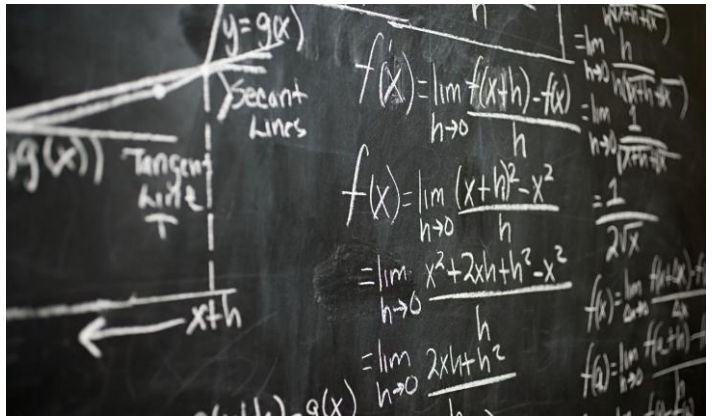
1. The ongoing interest on exposing the predictive components in returns,
2. the necessity for absolute accuracy and reliability in forecasting
3. the impressive advancement in computing power
4. data availability

Artificial Intelligence (AI) approaches provide:

(a) the *flexibility of working on nonlinear data* driven modelling whose forecasting properties are rather appealing - proven to act as universal approximators, able to fit any underlying data generating process (Hornik, Stinchcombe and White, 1989; Hornik, 1991),

(b) empirically supported that they can *predict both linear and nonlinear time series* (Zhang, 2001; Zhang et al., 2001).

Introduction-cont'd
Primary contribution



Our research efforts attempt to highlight benefits that can be attained in prediction and identify some key informative predictor variables, which under proper econometric modifications facilitate more trustworthy inspection into economic mechanisms of financial time series forecasting.

LSTM (Long Short-Term Memory) – why considered a "black box"?

choice of input features (several technical indicators, news, macroeconomic variables, stress and volatility indexes, prices, returns)

GOAL setting:

1. One of the goals in this study is to compare the effectiveness of the most well-known predictors in conventional finance, namely dividend-price (dp) and earnings-price (ep) and their derivatives (the cyclically-adjusted respective ratios, i.e. cadp and caep and their total return versions, i.e. trcadp and trcaep) in the context of LSTM networks. To our knowledge, while AI techniques have been used for forecasting in finance, they have not fully incorporated certain crucial financial variables as potential predictors.
2. We assume the presence of long-run equilibrium relationships between the variables and by exposing this drift differential, we construct alternative versions of the employed predictors which are anticipated to offer enhanced predictive gains.

Introduction-cont'd
Research planning

1. propose proper modifications to the classical dp and ep, by cyclically adjusting dividend and earnings as a moving average of the last 10 years following the rationale of Campbell and Shiller (1988) in the construction of their Cyclically-Adjusted Price-Earnings (CAPM)-therefore, we construct the cyclically adjusted dividend-price (cadp) and the cyclically-adjusted earnings-price (caep) ratios, as well as the total return cadp and caep (trcadp and trcaep respectively),
2. proceed on comparing the predictive performance of the cyclically adjusted ratios to their simple versions,
3. by identifying cointegration relationships within the basic ratios, we aim to econometrically modify the construction of both simple and cyclically adjusted ratios. *Our hypothesis is that these modified ratios will exhibit superior forecastability and deliver enhanced forecasting quality compared to the basic ratios*. This improvement is anticipated not only for return forecasts but also for the growth rates of dividends and earnings;
4. examine all included ratios' predictive performance both in-sample and out-of-sample and
5. by utilizing neural network techniques, we intend to forecast returns by using both the conventional ratios and their cyclically adjusted and modified counterparts.

Introduction-cont'd
Practical implications

- ✓ Our work is strongly related to the daily challenges faced by financial analysts, portfolio and risk managers, investors but also fellow researchers in the field.

- ✓ We propose the use of predictors whose construction is simple and straightforward and thus, addresses several concerns (mainly by practitioners) regarding the practicality of advanced econometric tools. Also, our modified predictors manage to tackle certain econometric issues (such as the sample bias, the forward-looking bias, stationarity, spurious regressions that the simple predictors have been accused of).

- ✓ The AI approach that we follow takes the analysis on the next level with enhanced predictive benefits but also even more reliable and robust results. The dominance of such techniques in the near future seems inevitable due to their flexibility, speed of adjustment, practicality and increased accuracy.

- ✓ *The proposed analysis provides an alternative look on predictors' construction either through AI or the traditional econometrics, that improves the quality but also quantity of our forecasts and helps us understand the full potential of machine learning techniques in time series forecasting by directly comparing outcomes as derived by both research routes.*

Lit. review at a glance- some key references

Just on S&P 500 market index:

- Siami-Namini and Namin (2018): LSTM vs ARIMA
- Sharma et al. (2021): LSTM vs ARIMAX
- Xiong et al. (2015): LSTM and volatility estimates
- Qiu et al. (2020): LSTM and "attention mechanism"
- Krauss et al. (2017): gradient boosted trees, deep neural networks, and random forests
- Fischer and Krauss (2018): LSTM
- Hossain et al. (2018): LSTM-GRU
- Kamalov et al. (2021): LSTM, random forest, Multilayer Perceptron and Logistic regression

And the list goes on…

Data & software

✓ **Data:** S&P 500 index, as data is available in Shiller's online data library.

Our choice is primarily motivated by market efficiency, data availability, computational feasibility and liquidity.

S&P 500 represents the leading gear of the U.S. stock market, representing a total 80% of available market capitalization (S&P Dow Jones Indices, 2017). Therefore, the index constitutes an overall sub-total of high liquidity and acts as an acid test for any trading strategy, attracting extensive investor scrutiny and analytics.

✓ Sample: 1926:01-2020:12

✓ **Software:** Eviews 14, Matlab R2016a and Python 3.5 (Python Software Foundation, 2016).

Methodological steps

1. Retrieve long-run equilibrium relationships in the examined vectors
2. Construct the modified versions of the predictors
3. Exploit predictors' dynamics on an LSTM framework of prediction through proper hyperparameter setting.
4. Include an in and out-of-sample regression analysis as dictated by traditional empirical analysis and
5. Compare outcomes with the LSTM network's forecastability so as to expose the true predictive benefits that lie in machine learning techniques.

Step 1

Johansen's cointegration approach

1. Assume a vector $w_t = [d_t\ p_t\ e_t]$.
2. Start with a VAR model

$$\boldsymbol{w_t = A_1 w_{t-1} + \cdots + A_k w_{t-k} + u_t,} \qquad \boldsymbol{u_t \sim IN(0, \Sigma)}$$

where w_t is (nx1) matrix and each of the A_i represents an (nxn) matrix of parameters.

3. Transform into a VECM

$$\boldsymbol{\Delta w_t = \Gamma_1 \Delta z_{t-1} + \cdots + \Gamma_\kappa \Delta z_{t-k+1} + \Pi z_{t-k} + u_t}$$

where $\Gamma_i = -(I - A_1 - \cdots - A_i)$ and $\Pi = -(\mathrm{I} - A_1 - \cdots - A_k)$.

4. Decide on the appropriate lag length following HQ criterion.
5. Consider linear deterministic trends in the vectors.
6. Test for uniqueness by posing restrictions to the parameters.

Cointegration test for the [$cad_t\ p_t$] vector and the null hypothesis of [1 -1]. Panel A includes findings on the Johansen technique assuming a linear deterministic trend in the data. Panel B entails the results for the imposed restriction [1 -1] examining the uniqueness of the vector. Data covers the period 1926:01-2020:12.

Panel A	#Coint. Vec.	Trace test stat.	5% critical value	
	0	26.64	0	15.49
	≤1	0.70	≤1	3.84
Panel B	H_0: [1 − 1]	**χ²-stat.**		
		11.83		

Cointegration test for the [$cae_t\ p_t$] vector and the null hypothesis of [1 -1]. Panel A includes findings on the Johansen technique assuming a linear deterministic trend in the data. Panel B entails the results for the imposed restriction [1 -1] examining the uniqueness of the vector. Data covers the period 1926:01-2020:12.

Panel A	#Coint. Vec.	Trace test stat.	5% critical value	
	0	16.77	0	15.94
	≤1	1.54	≤1	3.84
Panel B	H_0: [1 − 1]	**χ²-stat.**		
		1.28		

Cointegration test for the [$trcad_t\ trp_t$] vector and the null hypothesis of [1 -1]. Panel A includes findings on the Johansen technique assuming a linear deterministic trend in the data. Panel B entails the results for the imposed restriction [1 -1] examining the uniqueness of the vector. Data covers the period 1926:01-2020:12.

Panel A	#Coint. Vec.	Trace test stat.	5% critical value	
	0	15.98	0	15.49
	≤1	0.00	≤1	3.84
Panel B	H_0: [1 − 1]	**χ2-stat.**		
		9.46		

Cointegration test for the [$trcae_t\ trp_t$] vector and the null hypothesis of [1 -1]. Panel A includes findings on the Johansen technique assuming a linear deterministic trend in the data. Panel B entails the results for the imposed restriction [1 -1] examining the uniqueness of the vector. Data covers the period 1926:01-2020:12.

Panel A	#Coint. Vec.	Trace test stat.	5% critical value	
	0	22.62	0	15.49
	≤1	0.18	≤1	3.84
Panel B	H_0: [1 − 1]	**χ2-stat.**		
		5.32		

Cointegration test for the [$d_t\ p_t$] vector and the null hyphtesis of [1 -1]. Panel A includes findings on the Johansen technique assuming a linear deterministic trend in the data. Panel B entails the results for the imposed restriction [1 -1] examining the uniqueness of the vector. Data covers the period 1926:01-2020:12.

Panel A	#Coint. Vec.	Trace test stat.	5% critical value	
	0	25.04	0	15.49
	≤1	2.17	≤1	3.84
Panel B	H_0: [1 − 1]	**χ²-stat.**		
		15.07		

Cointegration test for the [$e_t\ p_t$] vector and the null hypothesis of [1 -1]. Panel A includes findings on the Johansen technique assuming a linear deterministic trend in the data. Panel B entails the results for the imposed restriction [1 -1] examining the uniqueness of the vector. Data covers the period 1926:01-2020:12.

Panel A	#Coint. Vec.	Trace test stat.	5% critical value	
	0	26.72	0	15.49
	≤1	0.80	≤1	3.84
Panel B	H_0: [1 − 1]	**χ²-stat.**		
		9.92		

Step 2

$$mcadp_t = cad_t - 0.610338p_t$$

$$mcaep_t = cae_t - 0.852164p_t$$

$$mtrcadp_t = trcad_t - 0.869654trp_t$$

$$mtrcaep_t = trcae_t - 0.928088trp_t$$

$$mdp_t = d_t - 0.820320p_t$$

$$mep_t = e_t - 0.887140p_t$$

Step 3

In-sample predictive regressions

Formulate continuously compounded nominal returns (r_t), excess returns (re_t) and real returns (rr_t) for h=36-, 60- and 84-months.

$$\boldsymbol{y_t(h) = \alpha + \beta x_t + u_t(h)}$$

where x_t is the predictor and y_t is either r_t, re_t, rr_t

Step 4

Out-of-sample (oos) predictive performance

We follow the typical oos coefficient of determination via the R^2_{oos} stat. as introduced by Campbell and Thompson (2008).

$$\boldsymbol{R^2_{oos} = 1 - [\sum_{k=1}^{t}(r_{t+k} - \hat{r}_{t+k})^2 / \sum_{k=1}^{t}(r_{t+k} - \bar{r}_{t+k})^2]}$$

Divide the sample into the training (1926:01-1956:12) and the test (till 2020:12) period - statistically imperative to have enough data so as to ensure reliability of the oos estimators.

2 techniques:

✓*the recursive*

✓*the full-sample*

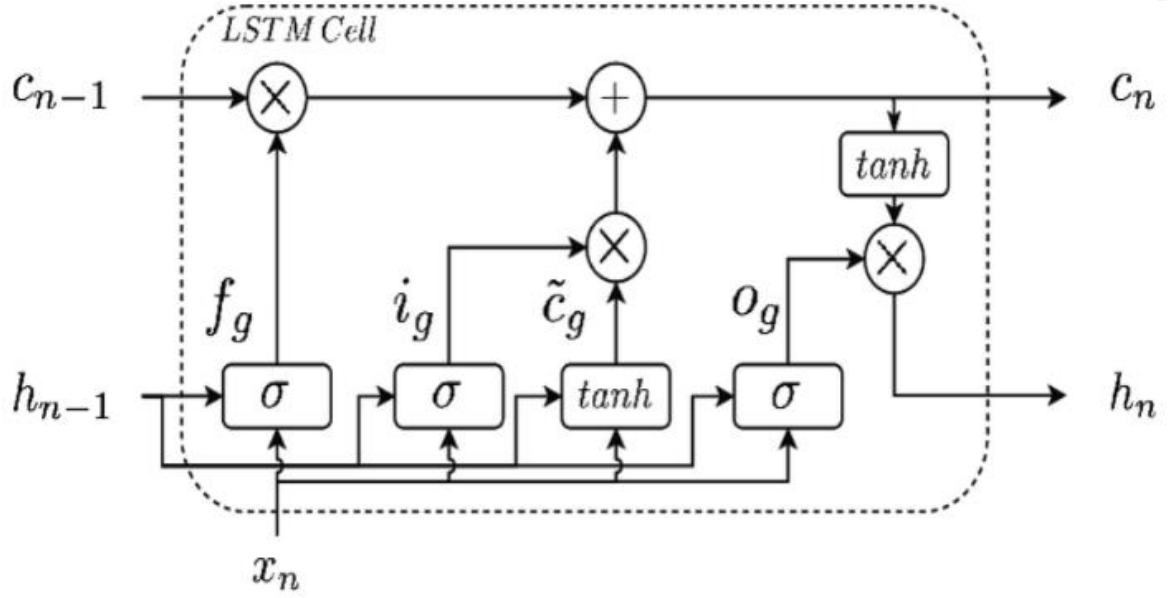
Step 5

LSTM networks – 3 things to bear in mind

1. How it works (memory cell visualization)
2. The mathematics
3. Hyperparameter setting

Step 5
1. Memory cell visualization – basic structure

Forget	Outputs a number between 0 and 1 (1 means “completely keep this” whereas, 0 implies “completely ignore this.”
Input	Chooses which new data need to be stored in the cell. First, a sigmoid layer, called the “input door layer” chooses which values will be modified. Next, a tanh layer makes a vector of new candidate values that could be added to the state.
Output	Decides what will be yield out of each cell. The yielded value will be based on the cell state along with the filtered and newly added data.



Step 5
2. The mathematics of it all

$$f_t = \sigma\left(W_f\left[h_{t-1,}x_t\right] + b_f\right)$$

$$i_t = \sigma\left(W_i\left[h_{t-1,}x_t\right] + b_i\right)$$

$$\tilde{c}_t = tanh(W_c\left[h_{t-1,}x_t\right] + b_c)$$

$$o_t = \sigma\left(W_o\left[h_{t-1,}x_t\right] + b_o\right)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$

$$h_t = o_t \circ tanh(c_t)$$

Step 5
3. Hyperparameter setting

LSTM architecture

Hyperparameter setting	
Input layer	100
LSTM layers	1 layer with h=25 hidden neurons
Dropout rate	0.1
Kernel and recurrent regularizer	L1 regularization with shrinkage 0.0001
Optimizer algorithm	Adam
Learning rate	0.0001 (Keras default learning rate)
Early stopping	1000 epochs, monitoring the validation loss
Validation split	0.2
Dense layer	1

Results - Summary stats

Correlation matrix and Descriptive Statistics. We show the summary statistics for nominal returns (r_t), excess returns (re_t) and real returns (rr_t), risk-free rate (rf_t), the cyclically-adjusted dp ($cadp_t$) and $mcadp_t$, total return cadp ($trcadp_t$) and $mtrcadp_t$, and the simple dp (dp_t) and mdp_t. The table depicts the correlation matrix between the series, the mean, standard deviation and the autocorrelation coefficient based on an AR(1) fitted model. Data is monthly, covering the period 1926:01-2020:12.

	r_t	re_t	rr_t	rf_t	$cadp_t$	$mcadp_t$	$trcadp_t$	$mtrcadp_t$	dp_t	mdp_t	Mean	Std	AR(1)
r_t	1										0.08	0.18	0.04
re_t	0.15	1									0.06	0.18	0.04
rr_t	0.99	0.13	1								0.05	0.18	0.02
rf_t	0.01	-0.99	0.02	1							0.03	0.03	0.89
$cadp_t$	0.01	-0.87	0.01	0.88	1						3.47	0.52	**0.93**
$mcadp_t$	-0.01	0.92	-0.02	-0.94	-0.89	1					-1.00	0.23	**0.85**
$trcadp_t$	0.02	-0.89	0.02	0.90	0.99	-0.91	1				3.66	0.46	**0.91**
$mtrcadp_t$	-0.01	0.94	-0.03	-0.95	-0.86	0.98	-0.89	1			-2.18	0.29	**0.60**
dp_t	0.01	-0.23	-0.00	0.23	0.63	-0.26	0.60	-0.18	1		-1.59	1.30	**0.90**
mdp_t	-0.03	0.80	-0.03	-0.82	-0.65	0.75	-0.67	0.78	0.04	1	-2.52	0.24	**0.54**

Results - Summary stats

Correlation matrix and Descriptive Statistics. We show the descriptive statistics for annual nominal returns (r_t), excess returns (re_t) and real returns (rr_t), risk-free rate (rf_t), the cyclically-adjusted ep ($caep_t$) and $mcaep_t$, total return caep ($trcaep_t$) and $mtrcaep_t$, and the simple ep (ep_t) and mep_t. The table depicts the correlation matrix between the series, the mean, standard deviation and the autocorrelation coefficient based on an AR(1) fitted model. Data is monthly, covering the period 1926:01-2020:12.

	r_t	re_t	rr_t	rf_t	$caep_t$	$mcaep_t$	$trcaep_t$	$mtrcaep_t$	ep_t	mep_t	Mean	Std	AR(1)
r_t	1										0.09	0.19	0.04
re_t	0.15	1									0.05	0.19	0.04
rr_t	0.99	0.13	1								0.06	0.19	0.02
rf_t	0.01	-0.99	0.02	1							0.03	0.03	0.89
$caep_t$	0.00	0.98	-0.01	-0.99	1						-2.83	0.40	**0.88**
$mcaep_t$	0.00	0.98	-0.01	-0.99	0.97	1					-1.89	0.31	**0.75**
$trcaep_t$	-0.01	0.99	-0.02	-0.99	0.99	0.99	1				-3.01	0.35	**0.90**
$mtrcaep_t$	-0.00	0.92	-0.02	-0.93	0.88	0.96	0.93	1			-2.02	0.30	**0.72**
ep_t	0.00	0.72	-0.01	-0.73	0.74	0.70	0.74	0.62	1		-2.75	0.42	**0.76**
mep_t	0.01	0.53	-0.01	-0.53	0.50	0.57	0.54	0.60	0.87	1	-2.21	0.35	**0.64**

Results cont'd In-sample predictability

In-sample predictability of nominal returns as derived by dividend-related predictors. Standard errors are GMM corrected. Data is monthly, covering the period 1926:01-2020:12.

h-months	*h=36*			*h=60*			*h=84*		
	b	t(b)	R^2	b	t(b)	R^2	b	t(b)	R^2
dp_t	0.06	1.27	**0.05**	0.09	1.40	**0.08**	0.09	1.01	**0.07**
mdp_t	0.66	7.02	**0.23**	1.03	9.12	**0.38**	1.14	13.67	**0.46**
$cadp_t$	-0.21	-2.67	0.10	-0.31	-3.16	0.14	-0.38	-2.86	**0.19**
$mcadp_t$	0.69	2.98	0.22	0.96	3.82	0.29	1.04	6.56	**0.32**
$trcadp_t$	-0.25	-2.84	0.11	-0.37	-3.41	0.16	-0.44	-3.20	**0.20**
$mtrcadp_t$	0.62	3.39	0.29	0.89	4.27	0.40	0.96	8.56	**0.44**

Results cont'd In-sample predictability

In-sample predictability of nominal returns including the earnings-related predictors. Standard errors are GMM corrected. Data is monthly, covering the period 1926:01-2020:12.

***h*-months**	*h=36*			*h=60*			*h=84*		
	b	**t(b)**	R^2	**b**	**t(b)**	R^2	**b**	**t(b)**	R^2
ep_t	0.24	2.87	**0.09**	0.34	2.50	0.12	0.49	2.90	**0.23**
mep_t	0.33	2.39	**0.12**	0.46	2.61	0.16	0.62	3.27	**0.28**
$caep_t$	0.39	3.35	**0.21**	0.56	4.79	0.28	0.66	6.87	**0.37**
$mcaep_t$	0.56	3.12	**0.27**	0.79	4.24	0.35	0.89	6.95	**0.44**
$trcaep_t$	0.46	3.26	**0.23**	0.67	4.39	0.32	0.77	7.10	**0.40**
$mtrcaep_t$	0.63	3.08	**0.32**	0.88	3.77	0.44	0.97	5.67	**0.50**

Results cont'd oos predictive performance

Out-of-sample (oos) forecasting. Oos forecasts for nominal returns as derived by the dividend-related predictors for the 12-,36-, 60-, 84-, 120-, and 144-months out. Data covers the period 1926:01-2020:12.

h-months	*h=36*			*h=60*			*h=84*		
	R^2	R^2_{oos} (rec)	R^2_{oos} (fs)	R^2	R^2_{oos} (rec)	R^2_{oos} (fs)	R^2	R^2_{oos} (rec)	R^2_{oos} (fs)
dp_t	0.05	-0.62		0.08	-0.33		0.07	-0.11	
mdp_t	0.23	**0.36**	**0.46**	0.38	0.40	0.50	0.46	**0.59**	**0.68**
$cadp_t$	0.10	-0.40		0.14	-0.98		0.19	-1.18	
$mcadp_t$	0.22	**0.23**	**0.31**	0.29	0.33	0.46	0.32	**0.45**	**0.54**
$trcadp_t$	0.11	-0.32		0.16	-0.12		0.20	-0.09	
$mtrcadp_t$	0.29	**0.41**	**0.61**	0.40	0.61	0.72	0.44	**0.73**	**0.80**

Results cont'd oos predictive performance

Out-of-sample (oos) forecasting. Oos forecasts for nominal returns as derived by the earnings-related predictors for the 12-,36-, 60-, 84-, 120-, and 144-months out. Data covers the period 1926:01-2020:12.

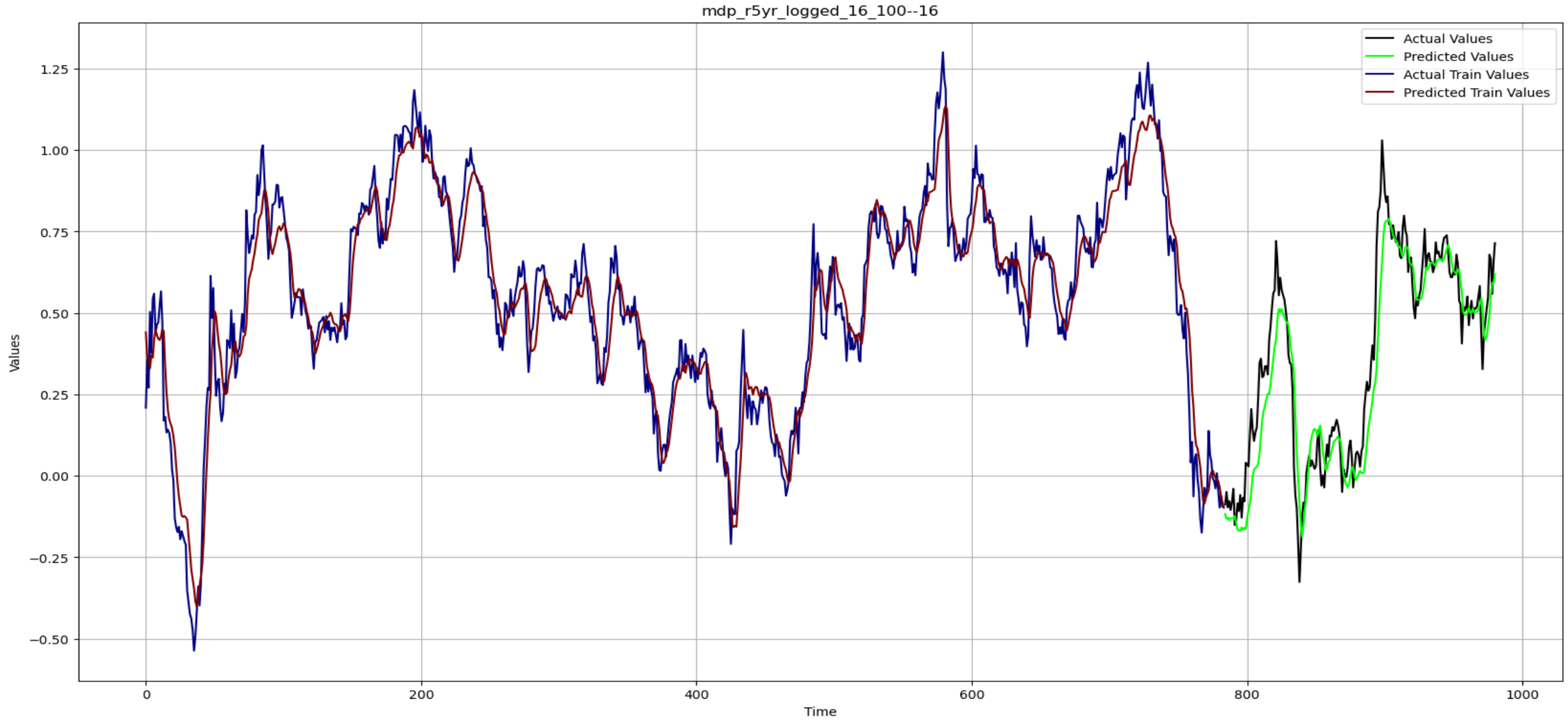
h-months	h=36			h=60			h=84		
	R^2	R^2_{oos} (rec)	R^2_{oos} (fs)	R^2	R^2_{oos} (rec)	R^2_{oos} (fs)	R^2	R^2_{oos} (rec)	R^2_{oos} (fs)
ep_t	0.09	-0.12		0.12	-0.08		0.23	0.08	
mep_t	0.12	**0.02**	**0.13**	0.16	0.04	0.22	0.28	**0.21**	**0.31**
$caep_t$	0.21	-1.12		0.28	-1.23		0.37	-2.99	
$mcaep_t$	0.27	**-0.86**	**0.16**	0.35	-0.34	0.32	0.44	**0.15**	**0.61**
$trcaep_t$	0.23	-0.16		0.32	-0.02		0.40	0.26	
$mtrcaep_t$	0.32	**0.03**	**0.12**	0.44	0.22	0.32	0.50	**0.34**	**0.51**

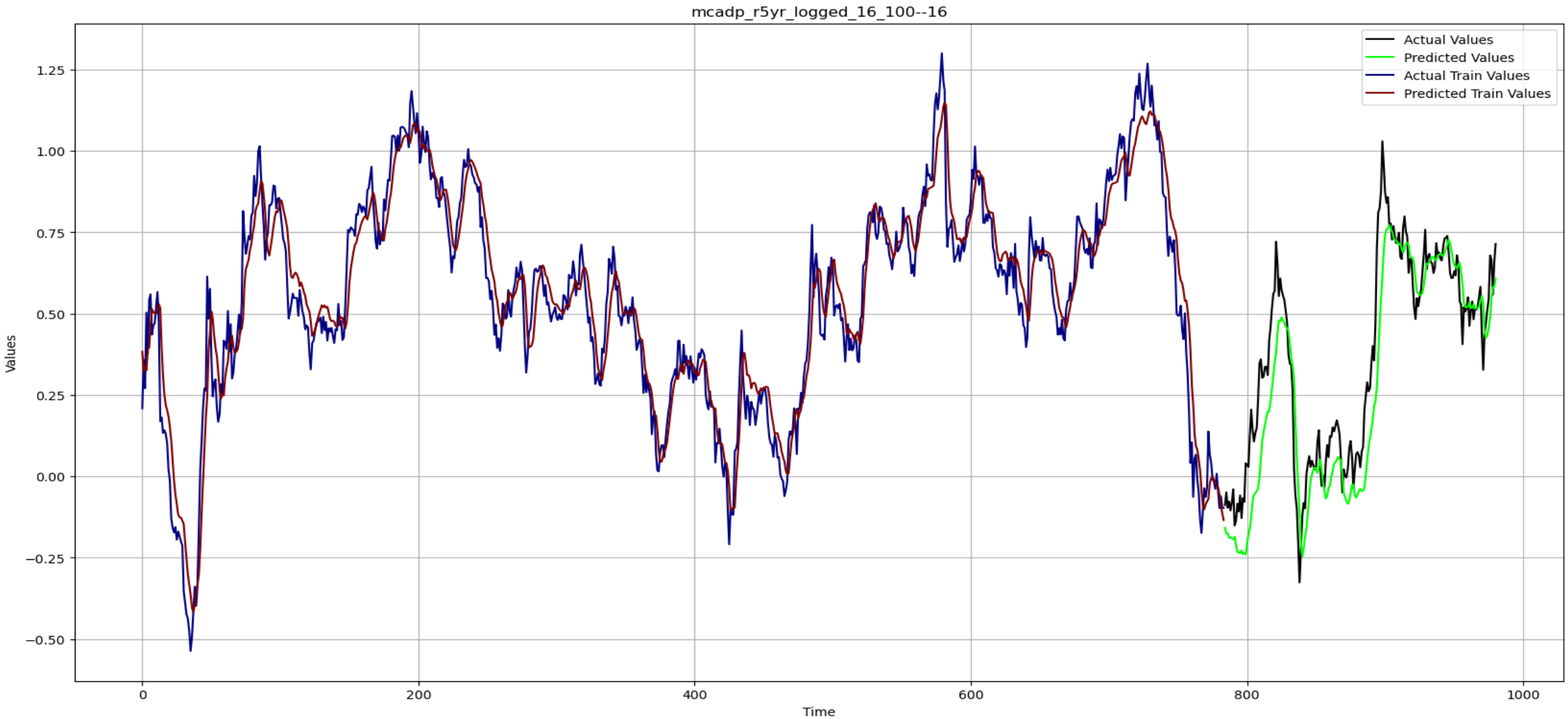
Results cont'd LSTM predictability findings

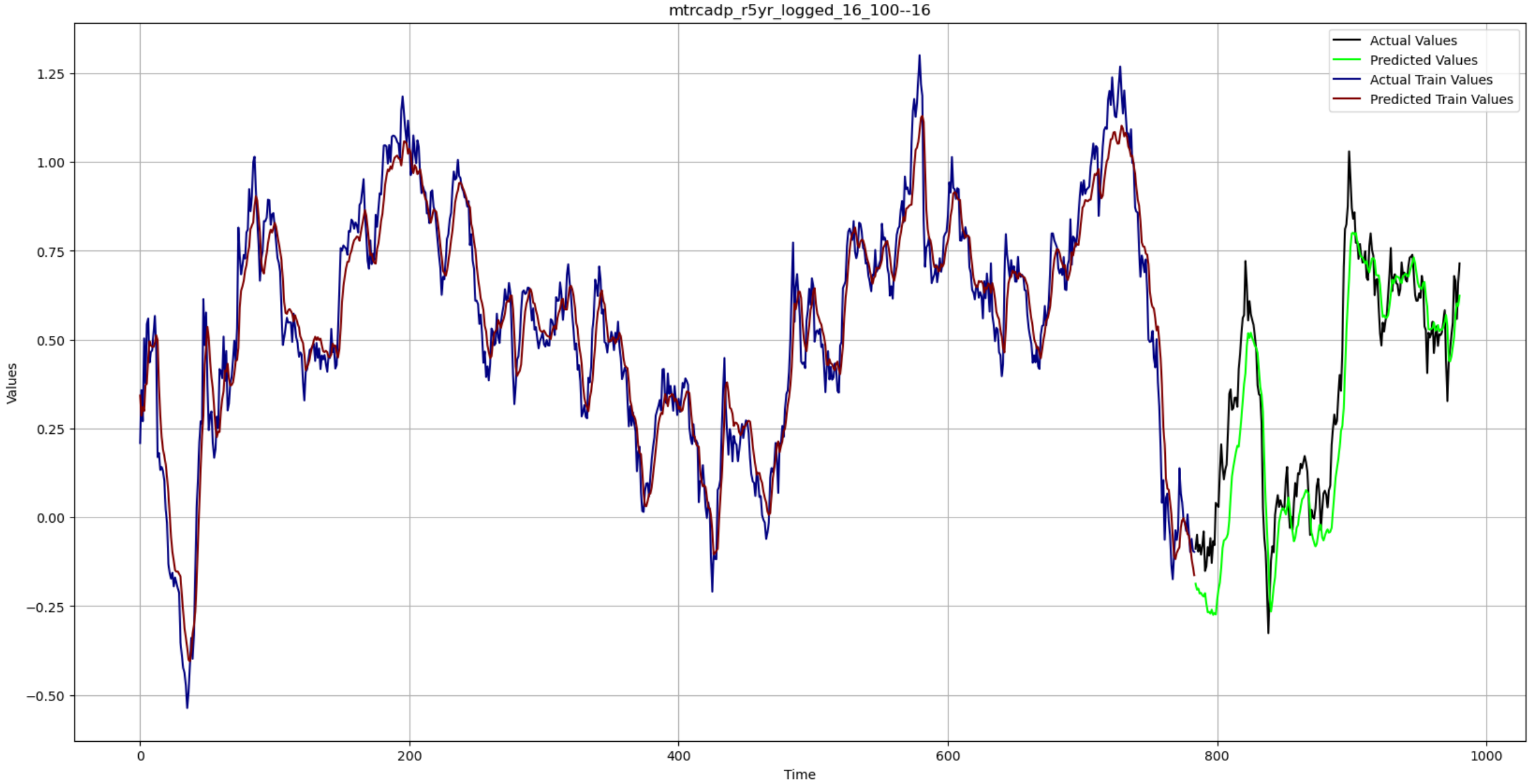
h-months	36		60		84	
	Train	Test	Train	Test	Train	Test
dp_t	0.496	**0.435**	0.700	**0.483**	0.793	**0.857**
mdp_t	0.588	**0.526**	0.767	**0.574**	0.879	**0.803**
$cadp_t$	0.517	**0.450**	0.711	**0.509**	0.824	**0.793**
$mcadp_t$	0.668	**0.673**	0.800	**0.676**	0.873	**0.838**
$trcadp_t$	0.469	**0.465**	0.748	**0.447**	0.765	**0.813**
$mtrcadp_t$	0.674	**0.628**	0.814	**0.675**	0.888	**0.781**

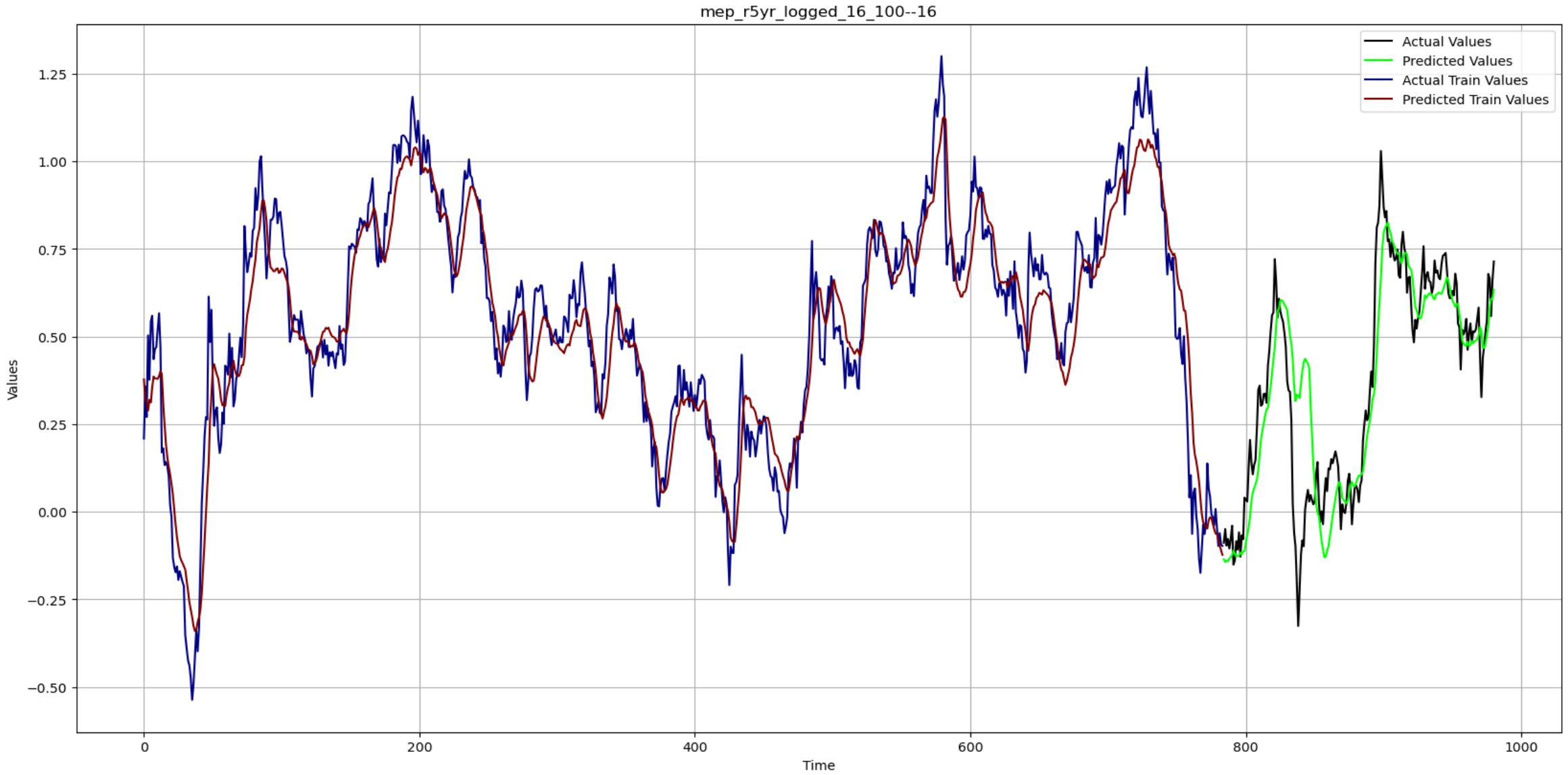
Results cont'd LSTM predictability findings

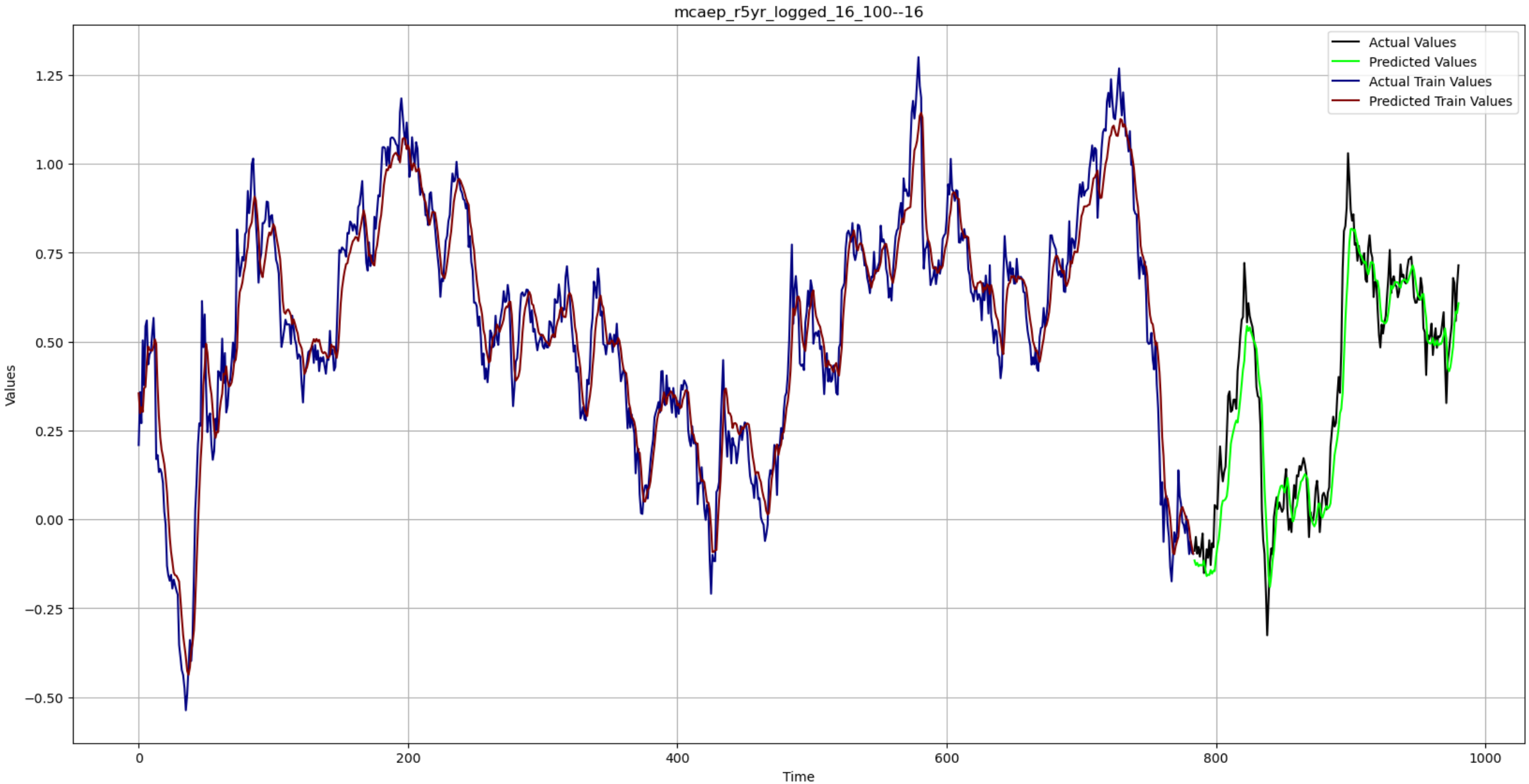
h-months	***36***		***60***		***84***	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*
ep_t	0.829	**0.687**	0.889	**0.654**	0.871	**0.624**
mep_t	0.823	**0.665**	0.911	**0.765**	0.941	**0.912**
$caep_t$	0.757	**0.691**	0.883	**0.810**	0.946	**0.921**
$mcaep_t$	0.853	**0.824**	0.930	**0.891**	0.941	**0.918**
$trcaep_t$	0.848	**0.810**	0.879	**0.814**	0.930	**0.850**
$mtrcaep_t$	0.843	**0.822**	0.915	**0.889**	0.946	**0.932**


mdp_r5yr_logged_16_100--16
Actual Values
Predicted Values
Actual Train Values
Predicted Train Values
Values
Time
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000


mcadp_r5yr_logged_16_100--16
Actual Values
Predicted Values
Actual Train Values
Predicted Train Values
Values
Time
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000


mtrcadp_r5yr_logged_16_100--16
Actual Values
Predicted Values
Actual Train Values
Predicted Train Values
Values
Time
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000


mep_r5yr_logged_16_100--16
Actual Values
Predicted Values
Actual Train Values
Predicted Train Values
Values
Time
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000


mcaep_r5yr_logged_16_100--16
Actual Values
Predicted Values
Actual Train Values
Predicted Train Values
Values
Time
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000


mtrcaep_r5yr_logged_16_100--16
Actual Values
Predicted Values
Actual Train Values
Predicted Train Values
Values
Time
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000

Robustness checks

1. ***Multivariate testing***
2. ***LSTM with no predictors***
3. ***Excess and real return predictability***

Concluding remarks

The utilization of machine learning algorithms for prediction purposes

has recently surfaced as a noteworthy research domain in the financial sector.

✓ ***Main findings***: *the modification on key-financial predictors leads to increased predictive benefits as proven by both the machine learning and the conventional econometric approaches.*

✓ ***Future work:***

- *Expand the data set to include extra international markets.*
- *Fix new predictors from ground up, and test similar hypotheses in return predictability.*

Thank you for your attention!