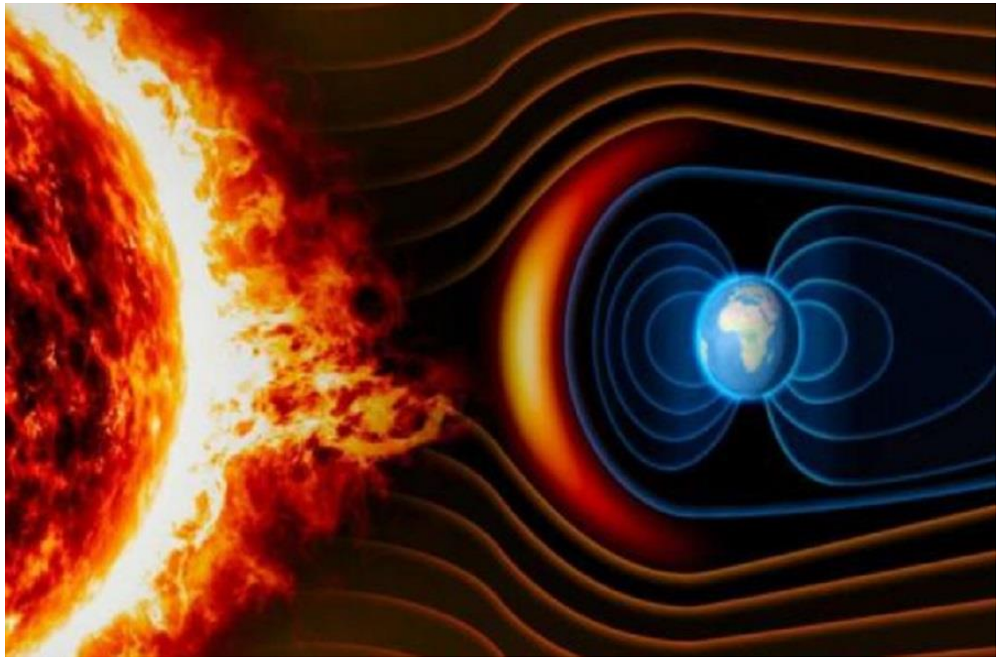
Prominence of data preparation in geomagnetic storm prediction using deep learning*

Marco Cristoforetti

*MC, R. Battiston, A. Gobbi, R. Iuppa, M. Piersanti, Scientific Reports (2022) 12:7631

Geomagnetic storms



Coronal Mass Ejections (CMEs) can eject billions of tons of coronal material and carry an embedded magnetic field, frozen in flux, that is stronger than the background solar wind interplanetary magnetic field (IMF) strength.

Geomagnetic storms

1989 Geomagnetic storm caused a **nine-hour outage of Hydro-Québec's electricity transmission system**

Scale	Description	Effect	Physical measure	Average Frequency (1 cycle = 11 years)
G 5	Extreme	**Power systems:** Widespread voltage control problems and protective system problems can occur, some grid systems may experience complete collapse or blackouts. Transformers may experience damage. **Spacecraft operations:** May experience extensive surface charging, problems with orientation, uplink/downlink and tracking satellites. **Other systems:** Pipeline currents can reach hundreds of amps, HF (high frequency) radio propagation may be impossible in many areas for one to two days, satellite navigation may be degraded for days, low-frequency radio navigation can be out for hours, and aurora has been seen as low as Florida and southern Texas (typically 40° geomagnetic lat.).	Kp = 9	4 per cycle (4 days per cycle)
G 4	Severe	**Power systems:** Possible widespread voltage control problems and some protective systems will mistakenly trip out key assets from the grid. **Spacecraft operations:** May experience surface charging and tracking problems, corrections may be needed for orientation problems. **Other systems:** Induced pipeline currents affect preventive measures, HF radio propagation sporadic, satellite navigation degraded for hours, low-frequency radio navigation disrupted, and aurora has been seen as low as Alabama and northern California (typically 45° geomagnetic lat.).	Kp = 8, including a 9-	100 per cycle (60 days per cycle)
G 3	Strong	**Power systems:** Voltage corrections may be required, false alarms triggered on some protection devices. **Spacecraft operations:** Surface charging may occur on satellite components, drag may increase on low-Earth-orbit satellites, and corrections may be needed for orientation problems. **Other systems:** Intermittent satellite navigation and low-frequency radio navigation problems may occur, HF radio may be intermittent, and aurora has been seen as low as Illinois and Oregon (typically 50° geomagnetic lat.).	Kp = 7	200 per cycle (130 days per cycle)
G 2	Moderate	**Power systems:** High-latitude power systems may experience voltage alarms, long-duration storms may cause transformer damage. **Spacecraft operations:** Corrective actions to orientation may be required by ground control; possible changes in drag affect orbit predictions. **Other systems:** HF radio propagation can fade at higher latitudes, and aurora has been seen as low as New York and Idaho (typically 55° geomagnetic lat.).	Kp = 6	600 per cycle (360 days per cycle)
G 1	Minor	**Power systems:** Weak power grid fluctuations can occur. **Spacecraft operations:** Minor impact on satellite operations possible. **Other systems:** Migratory animals are affected at this and higher levels; aurora is commonly visible at high latitudes (northern Michigan and Maine).	Kp = 5	1700 per cycle (900 days per cycle)

National Oceanic and Atmospheric Administration

ML - definitions

<u>Domain set:</u>

$\mathcal{X}$ Usually, you have vectors $\underline{x}$ of features

<u>Label set:</u>

$\mathcal{Y}$ for example $[0,1]$

There is a “correct” labeling function $f: \mathcal{X} \rightarrow \mathcal{Y}$

$$y_i = f(x_i)\ \forall i$$

We want the machine to LEARN a PREDICTOR

<u>Learner output:</u> a function $\boxed{\boldsymbol{h}: \boldsymbol{\mathcal{X}} \rightarrow \boldsymbol{\mathcal{Y}}}$

CENTER FOR DIGITAL INDUSTRY
FONDAZIONE BRUNO KESSLER

DSIP
DATA SCIENCE FOR INDUSTRY & PHYSICS

ML - definitions

<u>Training set:</u> $\boxed{\begin{array}{l} S = ((x_1, y_1), \dots, (x_N, y_N)) \\ \text{Sequence of pairs } \mathcal{X} \times \mathcal{Y} \end{array}}$ $=$ What the learner can use to determine the values of the predictor's parameters

S elements extracted using $\boldsymbol{\mathcal{D}}$, a probability distribution over the domain set $\boldsymbol{\mathcal{X}}$

We define the error of a predictor rule h to be:

$$L_{\mathcal{D},f}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x: h(x) \neq f(x)\})$$

The probability of randomly choosing an example x for which $h(x) \neq f(x)$

The learner is blind to the underlying $\boldsymbol{\mathcal{D}}$ and the function $\boldsymbol{f}$

TIFPA

deepPP

Empirical Risk Minimization

$L_{\mathcal{D},f}(h)$, the true error, is not directly available because we do not know $\mathcal{D}$ and f

TRAINING ERROR: error of the predictor on the training samples

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [N] : h(x_i) \neq y_i\}|}{N} \qquad [N] = \{1, \dots, N\}$$

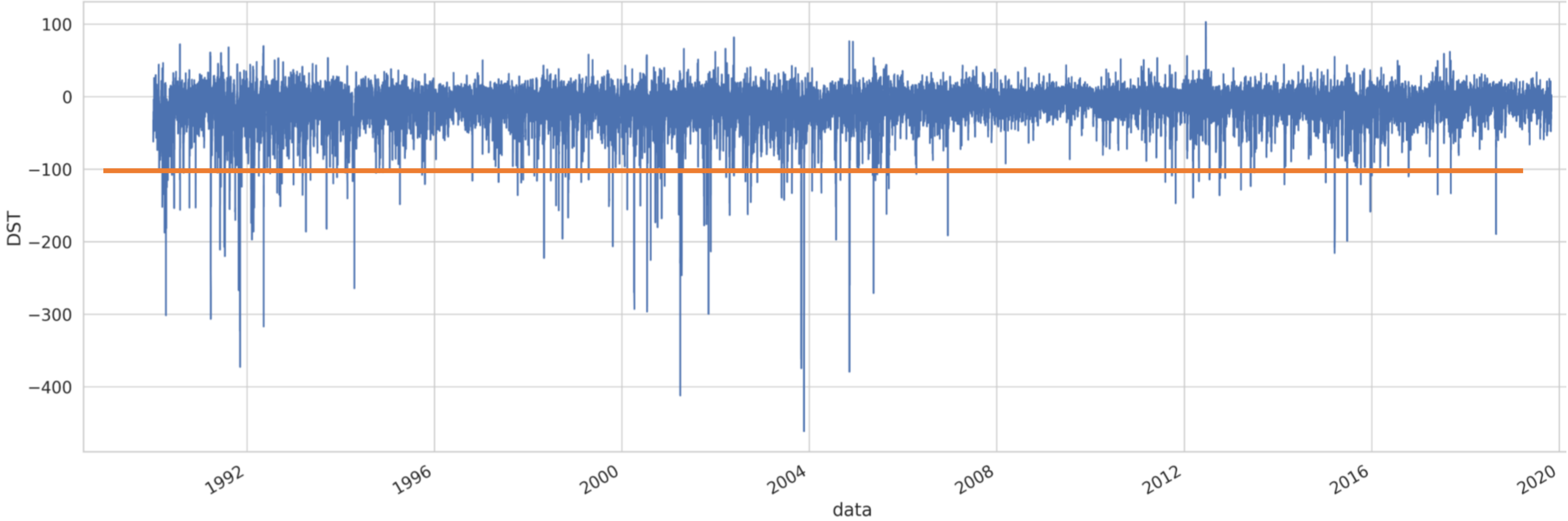
EMPIRICAL RISK/ERROR

HOW WE USE THE AVAILABLE DATA IS CRUCIAL

The Disturbance Storm Time Index

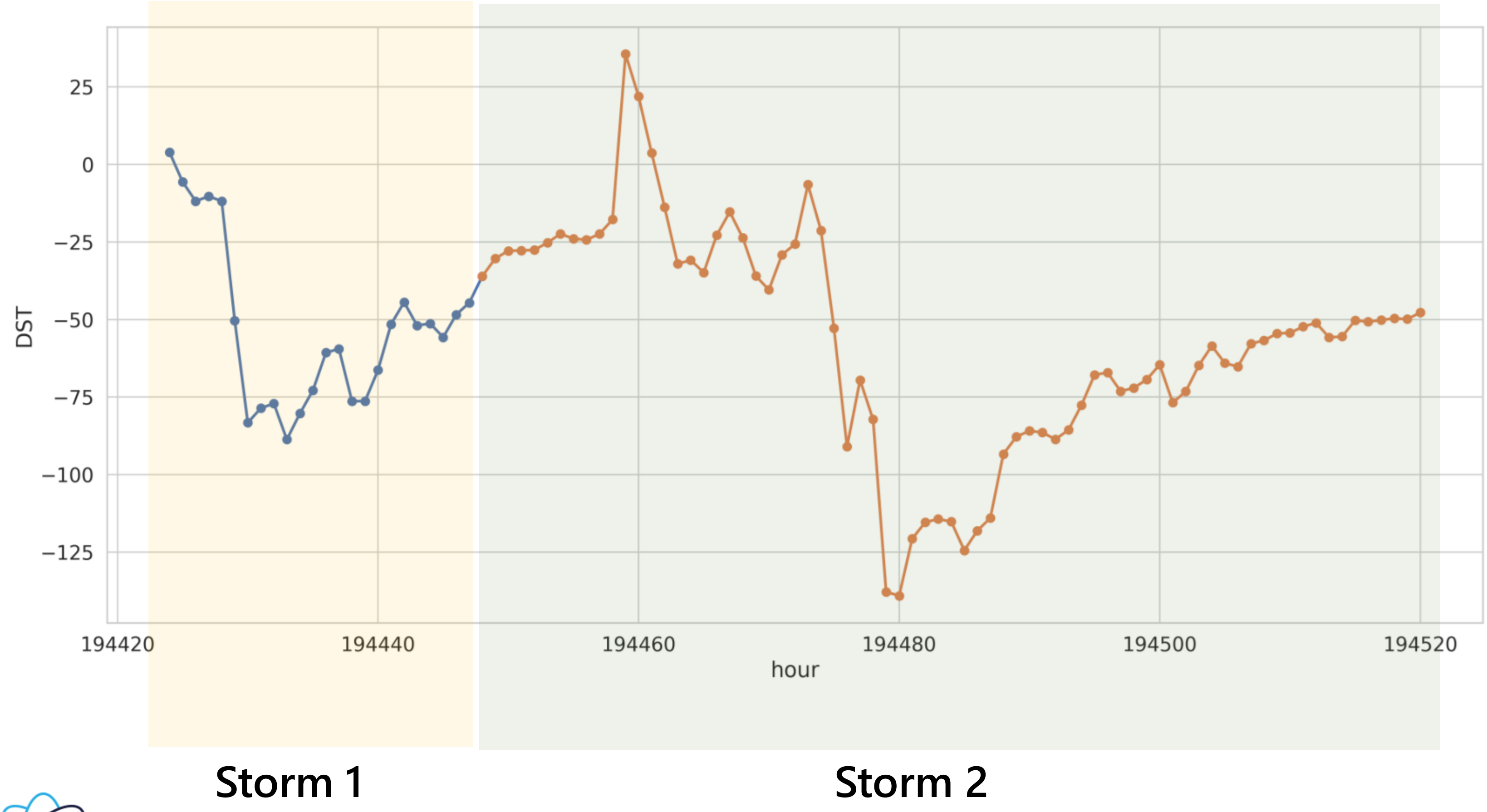
The Dst index is an index of magnetic activity derived from a network of near-equatorial geomagnetic observatories that measures the intensity of the globally symmetrical equatorial electrojet.
Dst is maintained at NGDC and is available via FTP from 1957 to the present.

DST > -20 nT	→ low
-20 nT ≥ DST > -50 nT	→ medium
-50 nT ≥ DST > -100 nT	→ high
DST ≤ -100 nT	→ intense



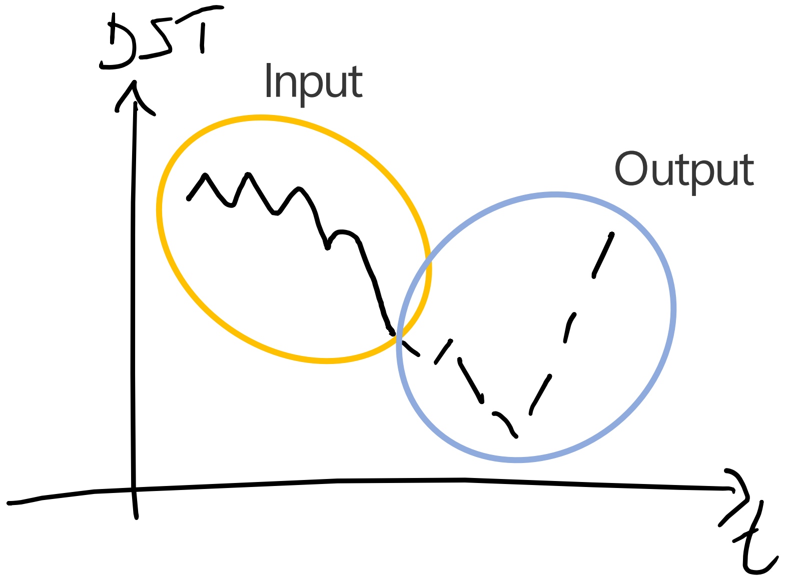
Forecast with Deep Learning

DST and geomagnetic storms



DST > -20 nT → low
-20 nT ≥ DST > -50 nT → medium
-50 nT ≥ DST > -100 nT → high
DST ≤ -100 nT → intense

Machine Learning for DST forecast

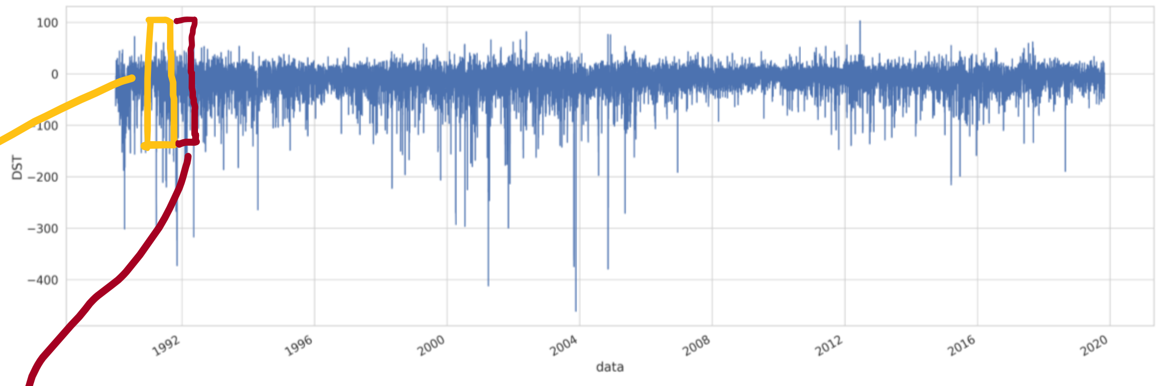


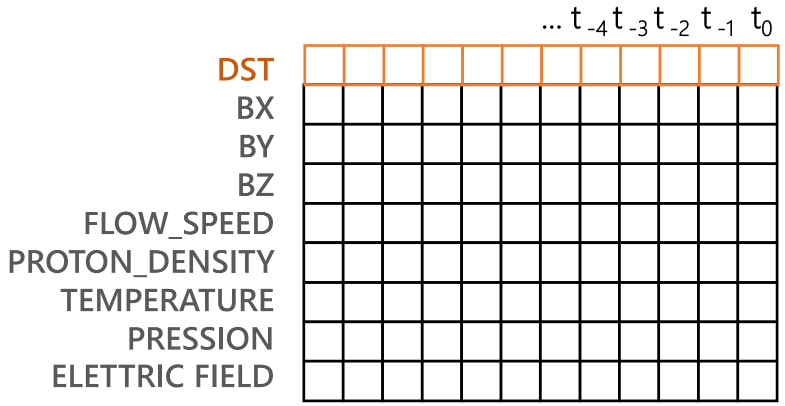
DST
BX, BY, BZ
FLOW_SPEED
PROTON_DENSITY
TEMPERATURE
PRESSURE
ELECTRIC FIELD

PRESSURE = α * PROTON_DENSITY*FLOW_SPEED^2
ELECTRIC = β * BZ * FLOW_SPEED

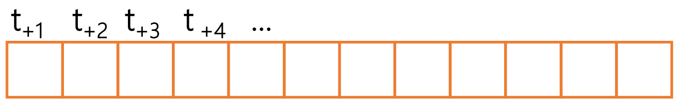
Input – output of the model

12 points (**hours**) to forecast the **next 12** points (**hours**)





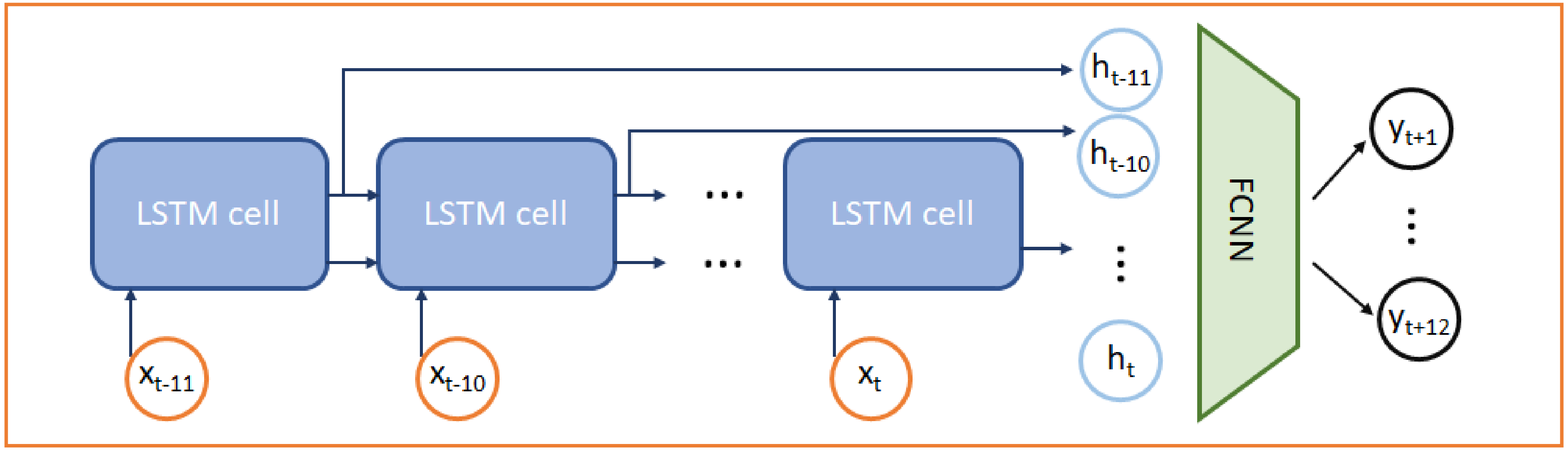
INPUT



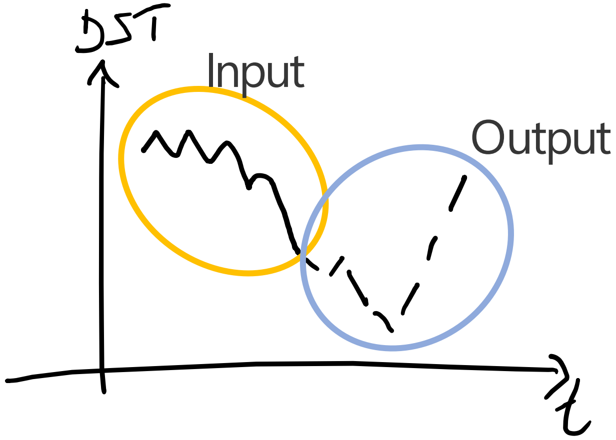
OUTPUT

DATASET
186 534 samples

Machine Learning for DST forecast



DST
BX, BY, BZ
FLOW_SPEED
PROTON_DENSITY
TEMPERATURE
PRESSURE
ELECTRIC FIELD

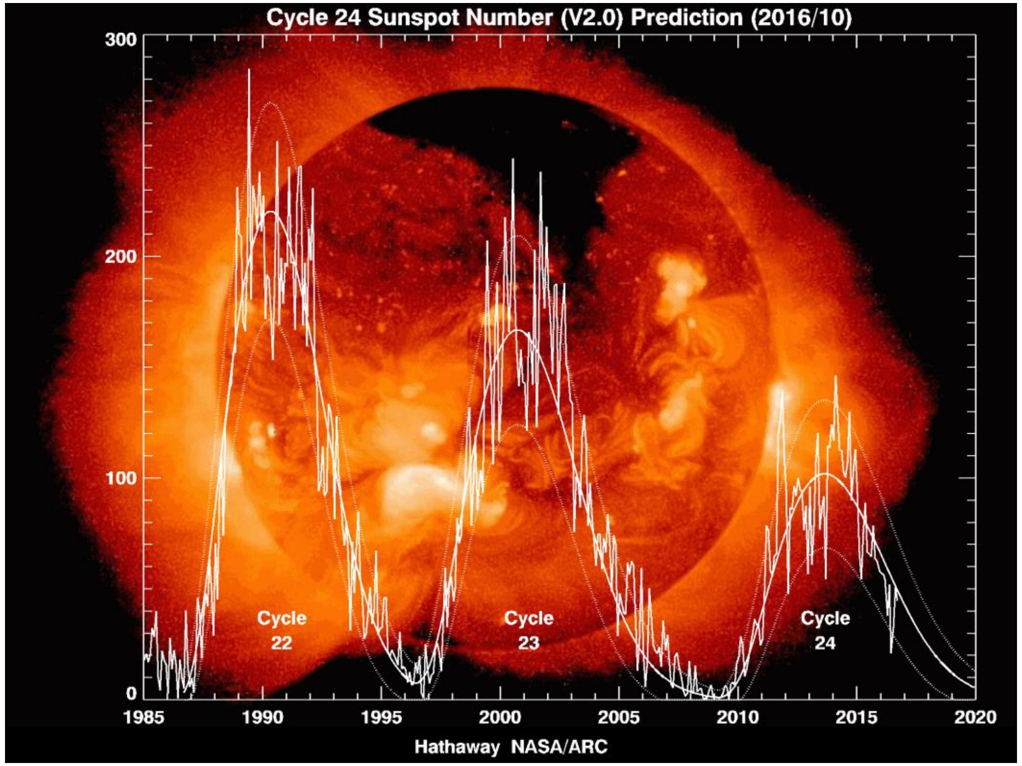


FONDAZIONE BRUNO KESSLER
CENTER FOR DIGITAL INDUSTRY

DSIP
DATA SCIENCE FOR INDUSTRY & PHYSICS

Dealing with time series and rare events

Periodicity and arrow of time. If data are periodic, it is safe to train the model considering at least one complete period and test it on different periods. being the arrow of time fixed and the future unknown, the training operation that make use of points that follow the data used in the test can introduce bias

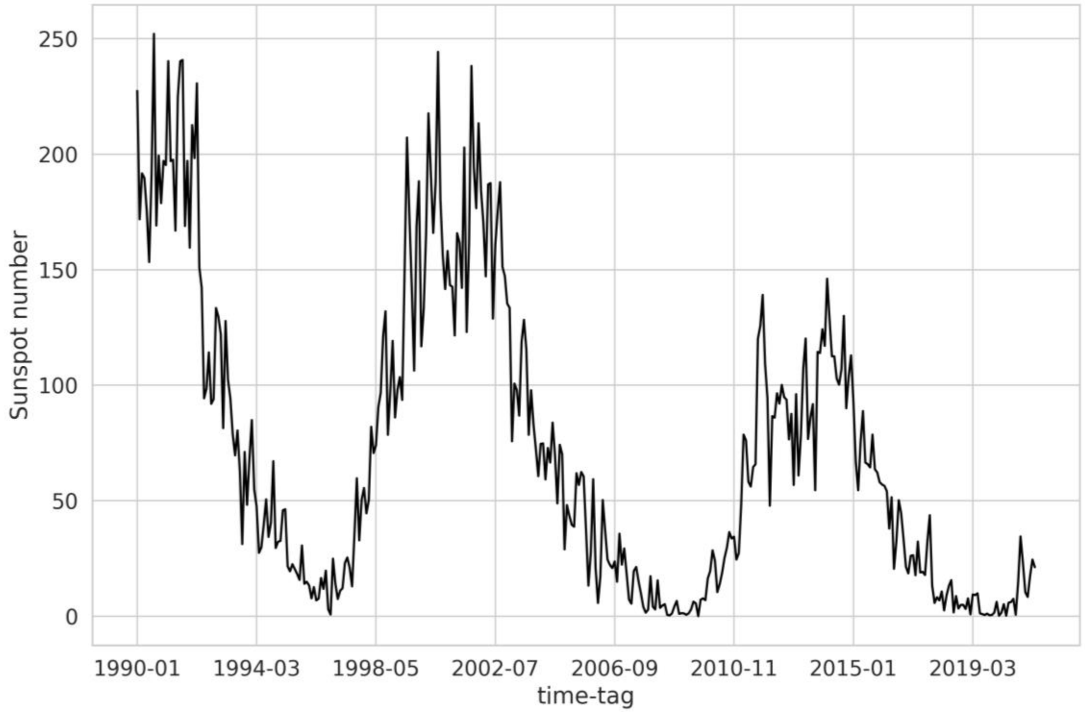


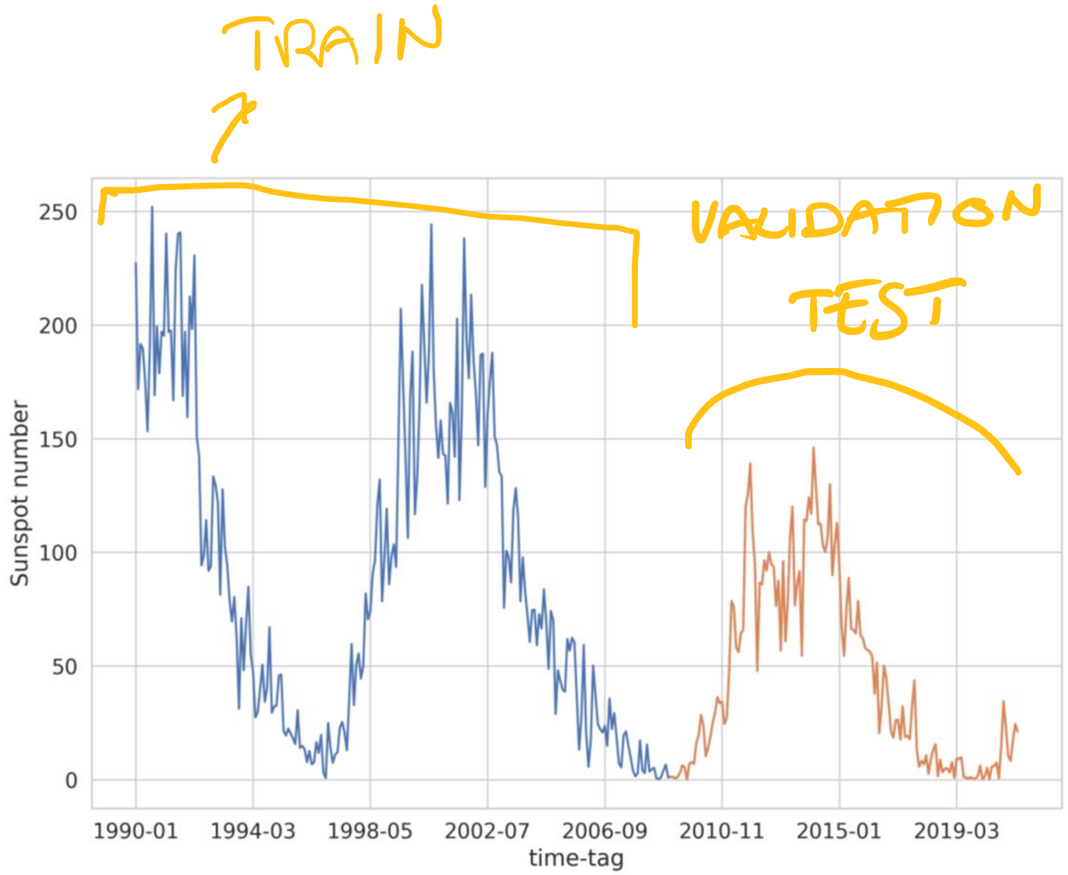
Forecast of rare events (storms). Training supervised DL model requires a **balanced sampling** of data referring to **quiet** and **storm** periods and **proper metrics** to measure the performances.

TIFPA

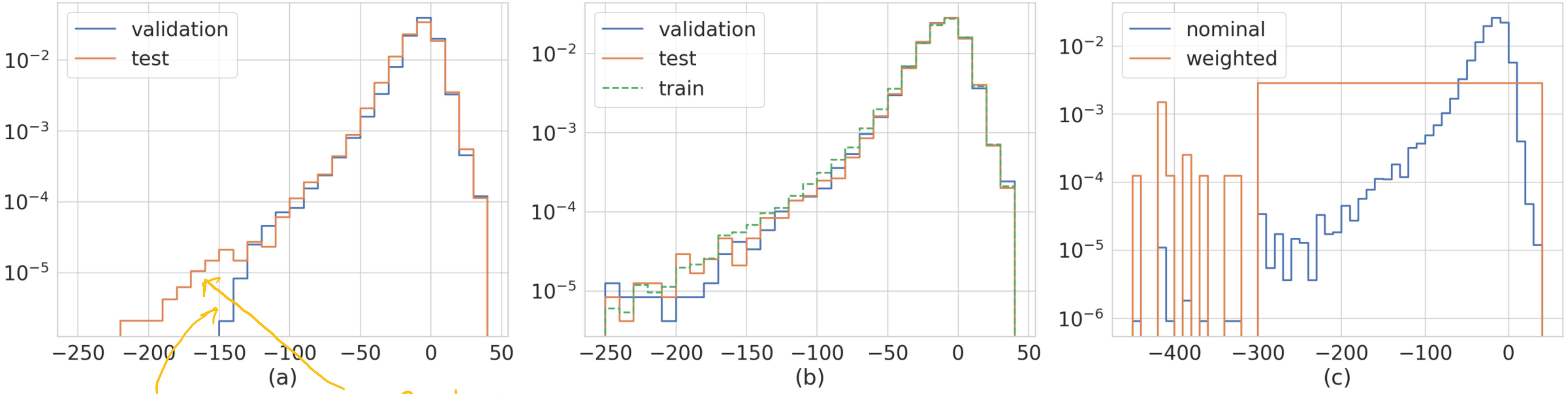
deeppp

Dataset preparation: solar activity



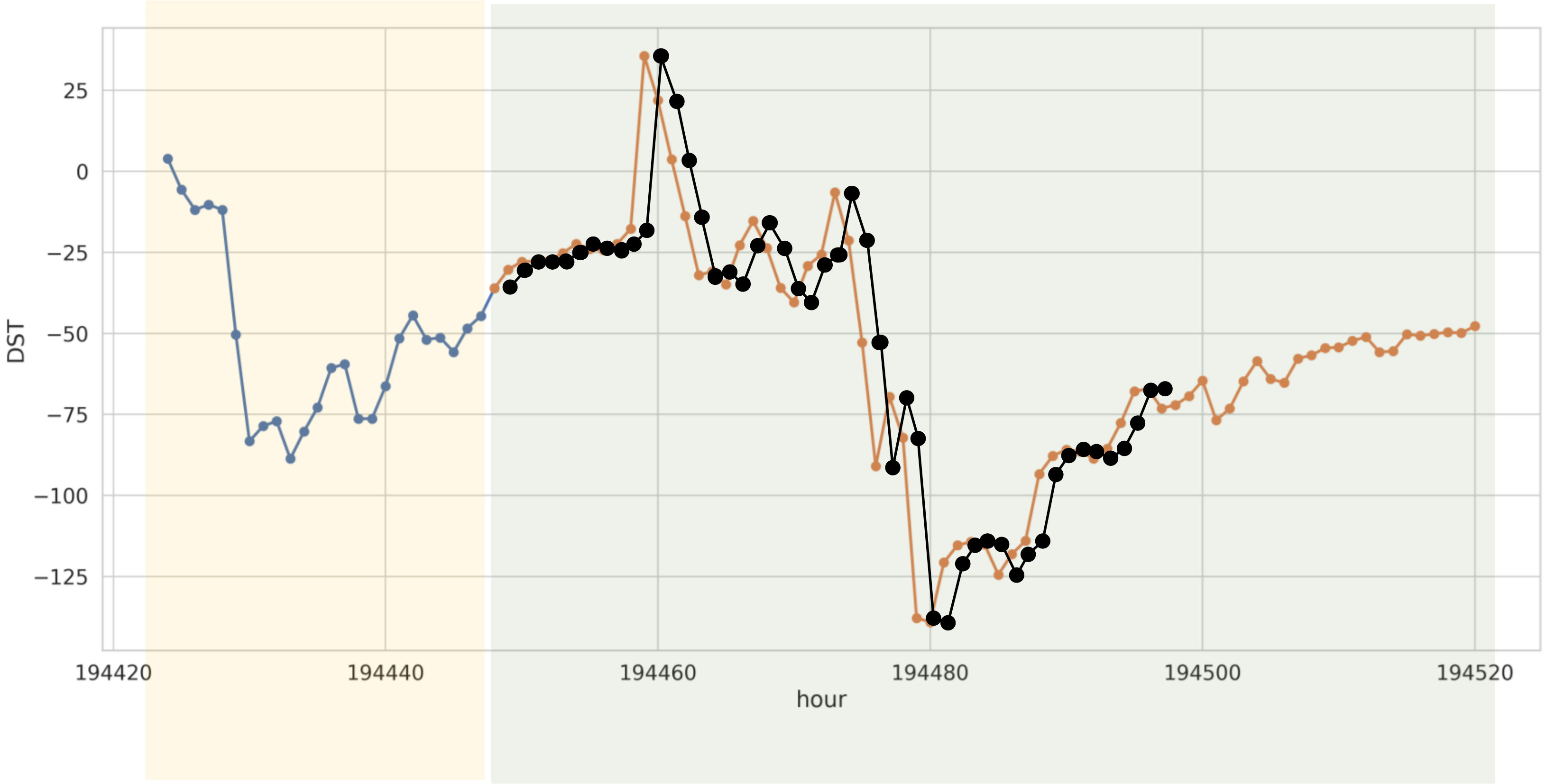


Dataset preparation
Random sampling and reweighting



1/2 cycle val   1/2 cycle test   RANDOM   TRAIN

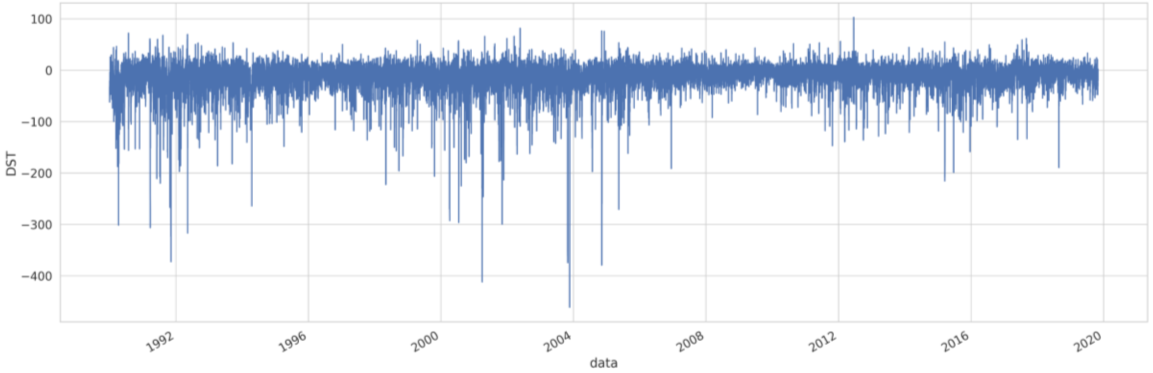
Persistence model



Storm 1 **Storm 2**

Looks good,
reproduce perfectly the shape,
but the algorithm
does not really learn

Results



Adopted metric: $$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_{pred_i} - y_{true}\right)^2}{N}}$$

Performance on the full dataset

	Train			Valid			Test		
	Pers	Nom	Weight	Pers	Nom	Weight	Pers	Nom	Weight
t + 1h	5.2	4.7	6.7	4.1	3.8	6.2	4.0	**3.8**	6.0
t + 2h	8.3	6.8	7.8	6.4	5.5	7.1	6.4	**5.4**	6.8
t + 3h	10.2	8.3	9.0	8.0	6.8	8.2	8.0	**6.7**	7.9
t + 4h	11.7	9.3	10.0	9.1	7.6	9.2	9.1	**7.5**	8.9
t + 5h	12.9	10.0	10.8	10.0	8.2	10.2	10.0	**8.1**	9.9
t + 6h	13.9	10.7	11.9	10.7	8.6	11.4	10.8	**8.7**	11.0
t + 7h	14.9	11.4	13.2	11.5	9.1	12.6	11.5	**9.2**	12.3
t + 8h	15.7	12.0	14.7	12.1	9.5	14.1	12.2	**9.7**	13.9
t + 9h	16.4	12.6	16.1	12.7	9.9	15.5	12.8	**10.1**	15.5
t + 10h	17.0	13.1	17.3	13.2	10.3	16.8	13.2	**10.4**	16.7
t + 11h	17.6	13.6	18.1	13.6	10.7	17.4	13.6	**10.8**	17.4
t + 12h	18.1	14.1	18.4	14.0	10.9	17.4	14.0	**11.0**	17.6

Using the weighted dataset the performances on the storms improve

	Dst> −20 nT			−20 nT > Dst > −50 nT			−50 nT > Dst > −100 nT			Dst< −100 nT		
	Pers	Nom	Weight	Pers	Nom	Weight	Pers	Nom	Weight	Pers	Full	Weight
t + 1h	**3.4**	**3.4**	5.8	5.6	**4.1**	6.2	10.0	**9.9**	11.8	**17.0**	18.2	23.6
t + 2h	5.2	**4.7**	6.6	8.8	**6.7**	7.8	17.5	14.5	**13.8**	31.4	28.6	**27.9**
t + 3h	6.3	**5.6**	7.5	11.0	**8.5**	9.1	22.9	18.6	**16.8**	43.2	40.0	**35.0**
t + 4h	7.0	**6.1**	8.5	12.5	**9.7**	10.4	27.4	21.5	**18.4**	54.1	49.8	**43.2**
t + 5h	7.5	**6.4**	9.2	13.7	**10.5**	11.7	31.0	24.4	**20.9**	63.8	56.8	**46.6**
t + 6h	7.9	**6.6**	10.2	14.9	**11.1**	13.0	34.3	26.9	**22.6**	73.2	63.5	**50.4**
t + 7h	8.2	**6.7**	11.2	15.9	**11.7**	14.7	37.6	29.7	**24.1**	81.4	68.3	**53.8**
t + 8h	8.5	**6.9**	12.7	16.8	**12.4**	16.3	40.5	32.1	**25.5**	87.9	73.0	**56.8**
t + 9h	8.8	**7.0**	14.0	17.6	**12.9**	17.7	42.9	34.3	**27.4**	93.4	77.3	**60.1**
t + 10h	9.0	**7.1**	15.3	18.2	**13.3**	18.7	44.9	36.1	**29.0**	97.4	82.4	**64.6**
t + 11h	9.2	**7.2**	15.9	18.9	**13.7**	19.4	46.7	37.8	**30.6**	100.7	86.7	**69.9**
t + 12h	9.4	**7.3**	15.9	19.5	**14.0**	19.9	48.1	39.3	**31.8**	103.4	90.4	**73.8**

Persistence by Dynamic Time Warping

Forecast horizon	Time shift						
	0 h	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**
$t + 1h$	0.003	**0.997**	0.0	0.0	0.0	0.0	0.0
$t + 2h$	0.003	0.003	**0.994**	0.0	0.0	0.0	0.0
$t + 3h$	0.004	0.003	0.003	**0.991**	0.0	0.0	0.0
$t + 4h$	0.003	0.003	0.003	0.003	**0.988**	0.0	0.0
$t + 5h$	0.004	0.003	0.003	0.003	0.003	**0.984**	0.0
$t + 6h$	0.004	0.003	0.003	0.003	0.003	0.003	**0.981**

Persistent

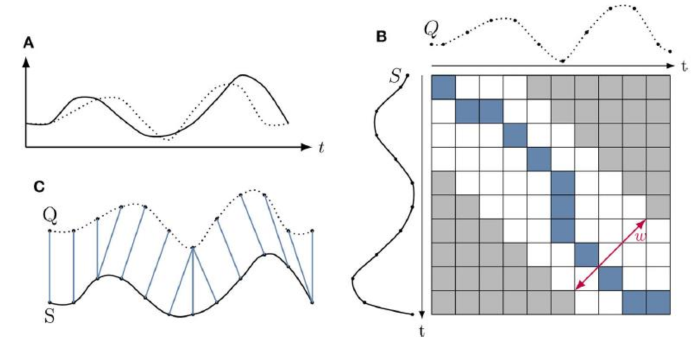


FIGURE 4 | (A) Shows two time series we want to compare. **(B)** Illustrates the cumulative distance matrix, together with a warping window *w* and the ideal warping path *P* in blue. **(C)** Illustrates the warping path *P* aligning the two time series.

Dynamic Time Warping as a New Evaluation for Dst Forecast With Machine Learning, B. Laperre, J.A. and G. Lapenta, Front. Astron. Space Sci., 22 July 2020

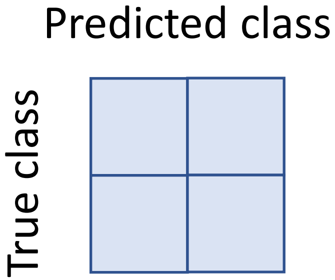
	0 h	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**
t + 1h	**0.37**	**0.37**	0.12	0.07	0.05	0.01	0.01
t + 2h	0.15	0.31	**0.33**	0.13	0.04	0.02	0.02
t + 3h	0.08	0.14	0.25	**0.31**	0.11	0.07	0.04
t + 4h	0.08	0.08	0.15	0.22	**0.26**	0.13	0.08
t + 5h	0.07	0.09	0.05	0.13	0.24	**0.26**	0.15
t + 6h	0.11	0.11	0.07	0.11	0.12	0.22	**0.27**

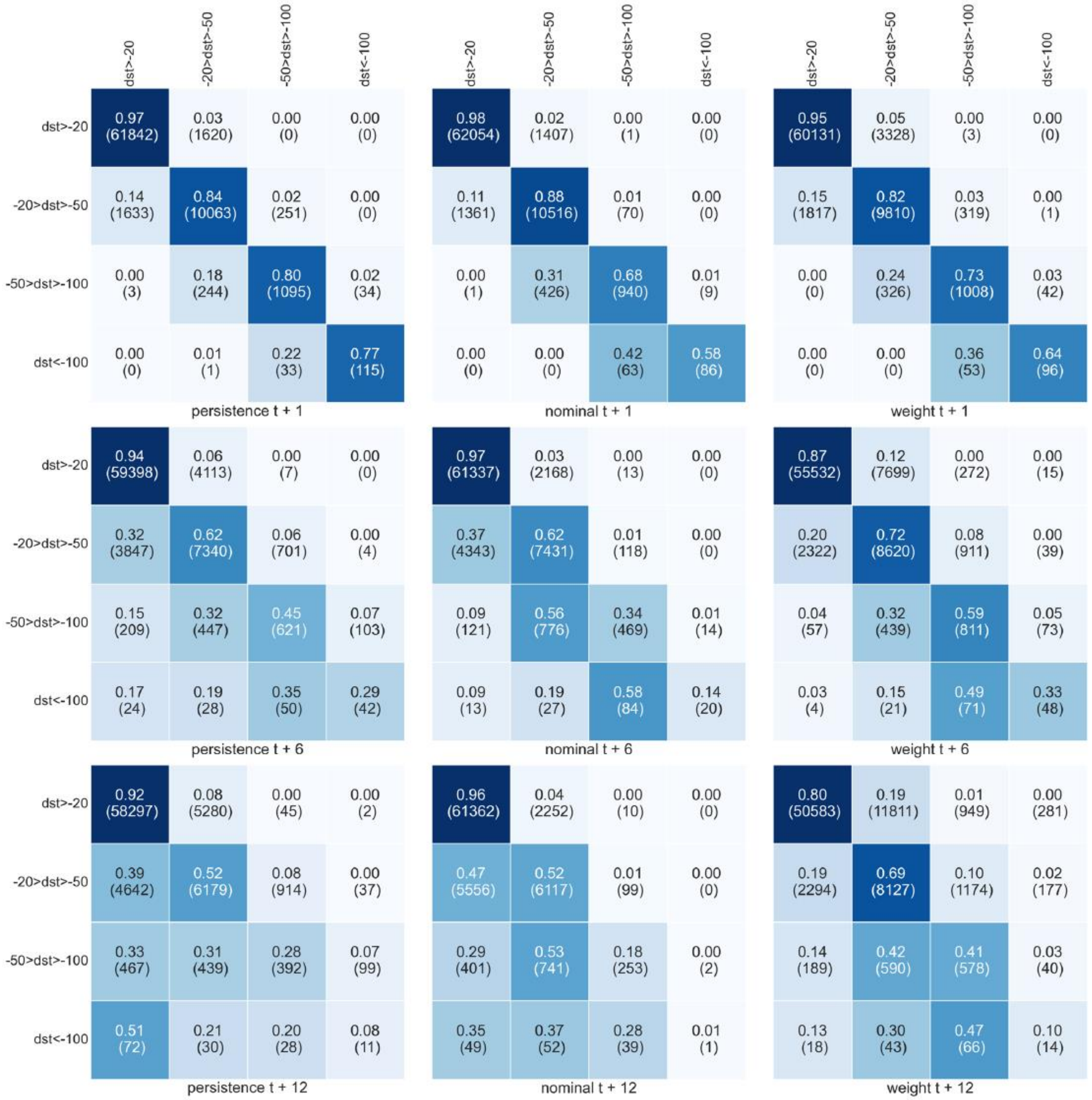
Nominal

	0 h	**1 h**	**2 h**	**3 h**	**4 h**	**5 h**	**6 h**
t + 1h	**0.34**	0.21	0.14	0.08	0.11	0.08	0.05
t + 2h	0.17	0.19	**0.27**	0.12	0.09	0.07	0.09
t + 3h	0.13	0.13	0.23	**0.17**	0.15	0.12	0.08
t + 4h	**0.18**	0.10	0.15	**0.18**	**0.18**	0.14	0.07
t + 5h	0.19	0.08	0.09	0.12	**0.20**	0.19	0.13
t + 6h	**0.23**	0.07	0.15	0.04	0.11	**0.23**	0.17

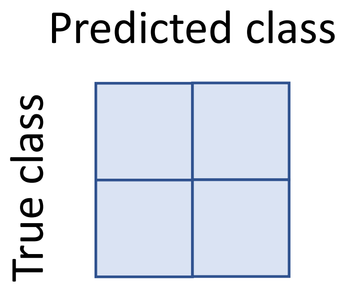
Weighted

Confusion matrix
Full dataset





Confusion matrix
Data with input
DST > -50nT



	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.95 (60245)	0.05 (3073)	0.00 (0)	0.00 (0)
-20>dst>-50	0.31 (3077)	0.69 (6747)	0.00 (0)	0.00 (0)
-50>dst>-100	0.22 (67)	0.78 (243)	0.00 (0)	0.00 (0)
dst<-100	0.36 (5)	0.64 (9)	0.00 (0)	0.00 (0)

persistence t + 3

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.97 (61313)	0.03 (1999)	0.00 (6)	0.00 (0)
-20>dst>-50	0.33 (3238)	0.67 (6563)	0.00 (23)	0.00 (0)
-50>dst>-100	0.08 (24)	0.75 (231)	0.17 (54)	0.00 (1)
dst<-100	0.14 (2)	0.57 (8)	0.29 (4)	0.00 (0)

nominal t + 3

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.92 (58207)	0.08 (5084)	0.00 (26)	0.00 (1)
-20>dst>-50	0.22 (2177)	0.76 (7447)	0.02 (199)	0.00 (1)
-50>dst>-100	0.06 (19)	0.59 (184)	0.33 (103)	0.01 (4)
dst<-100	0.07 (1)	0.50 (7)	0.36 (5)	0.07 (1)

weight t + 3

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.94 (59348)	0.06 (3931)	0.00 (0)	0.00 (0)
-20>dst>-50	0.40 (3815)	0.60 (5823)	0.00 (0)	0.00 (0)
-50>dst>-100	0.41 (207)	0.59 (293)	0.00 (0)	0.00 (0)
dst<-100	0.49 (24)	0.51 (25)	0.00 (0)	0.00 (0)

persistence t + 6

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.97 (61269)	0.03 (2000)	0.00 (10)	0.00 (0)
-20>dst>-50	0.45 (4318)	0.55 (5296)	0.00 (24)	0.00 (0)
-50>dst>-100	0.24 (121)	0.65 (325)	0.11 (54)	0.00 (0)
dst<-100	0.27 (13)	0.43 (21)	0.27 (13)	0.04 (2)

nominal t + 6

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.88 (55511)	0.12 (7498)	0.00 (255)	0.00 (15)
-20>dst>-50	0.24 (2314)	0.72 (6936)	0.04 (360)	0.00 (28)
-50>dst>-100	0.11 (57)	0.54 (268)	0.31 (154)	0.04 (21)
dst<-100	0.08 (4)	0.41 (20)	0.41 (20)	0.10 (5)

weight t + 6

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.92 (58243)	0.08 (4926)	0.00 (0)	0.00 (0)
-20>dst>-50	0.49 (4617)	0.51 (4798)	0.00 (0)	0.00 (0)
-50>dst>-100	0.59 (462)	0.41 (320)	0.00 (0)	0.00 (0)
dst<-100	0.72 (72)	0.28 (28)	0.00 (0)	0.00 (0)

persistence t + 12

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.97 (61266)	0.03 (1894)	0.00 (9)	0.00 (0)
-20>dst>-50	0.59 (5511)	0.41 (3869)	0.00 (35)	0.00 (0)
-50>dst>-100	0.51 (401)	0.43 (340)	0.05 (41)	0.00 (0)
dst<-100	0.49 (49)	0.41 (41)	0.09 (9)	0.01 (1)

nominal t + 12

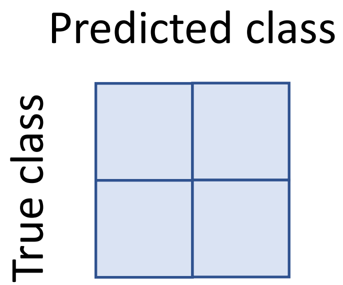
	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.80 (50577)	0.18 (11444)	0.01 (881)	0.00 (267)
-20>dst>-50	0.24 (2291)	0.67 (6343)	0.07 (639)	0.02 (142)
-50>dst>-100	0.24 (189)	0.49 (381)	0.24 (191)	0.03 (21)
dst<-100	0.18 (18)	0.38 (38)	0.36 (36)	0.08 (8)

weight t + 12

FBK CENTER FOR DIGITAL INDUSTRY
FONDAZIONE BRUNO KESSLER

DSIP DATA SCIENCE FOR INDUSTRY & PHYSICS

Confusion matrix Data with input DST > -20nT



persistence t + 3

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	1.00 (51499)	0.00 (0)	0.00 (0)	0.00 (0)
-20>dst>-50	1.00 (1380)	0.00 (0)	0.00 (0)	0.00 (0)
-50>dst>-100	1.00 (50)	0.00 (0)	0.00 (0)	0.00 (0)
dst<-100	1.00 (5)	0.00 (0)	0.00 (0)	0.00 (0)

nominal t + 3

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	1.00 (51357)	0.00 (140)	0.00 (2)	0.00 (0)
-20>dst>-50	0.85 (1175)	0.15 (204)	0.00 (1)	0.00 (0)
-50>dst>-100	0.38 (19)	0.50 (25)	0.12 (6)	0.00 (0)
dst<-100	0.40 (2)	0.60 (3)	0.00 (0)	0.00 (0)

weight t + 3

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.98 (50597)	0.02 (894)	0.00 (7)	0.00 (1)
-20>dst>-50	0.69 (954)	0.29 (404)	0.02 (22)	0.00 (0)
-50>dst>-100	0.32 (16)	0.50 (25)	0.16 (8)	0.02 (1)
dst<-100	0.20 (1)	0.80 (4)	0.00 (0)	0.00 (0)

persistence t + 6

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	1.00 (50808)	0.00 (0)	0.00 (0)	0.00 (0)
-20>dst>-50	1.00 (1941)	0.00 (0)	0.00 (0)	0.00 (0)
-50>dst>-100	1.00 (165)	0.00 (0)	0.00 (0)	0.00 (0)
dst<-100	1.00 (20)	0.00 (0)	0.00 (0)	0.00 (0)

nominal t + 6

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	1.00 (50646)	0.00 (159)	0.00 (3)	0.00 (0)
-20>dst>-50	0.90 (1744)	0.10 (194)	0.00 (3)	0.00 (0)
-50>dst>-100	0.58 (95)	0.35 (57)	0.08 (13)	0.00 (0)
dst<-100	0.50 (10)	0.40 (8)	0.10 (2)	0.00 (0)

weight t + 6

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.96 (48867)	0.04 (1824)	0.00 (110)	0.00 (7)
-20>dst>-50	0.66 (1276)	0.30 (583)	0.03 (65)	0.01 (17)
-50>dst>-100	0.33 (54)	0.39 (65)	0.24 (40)	0.04 (6)
dst<-100	0.10 (2)	0.65 (13)	0.15 (3)	0.10 (2)

persistence t + 12

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	1.00 (49853)	0.00 (0)	0.00 (0)	0.00 (0)
-20>dst>-50	1.00 (2641)	0.00 (0)	0.00 (0)	0.00 (0)
-50>dst>-100	1.00 (377)	0.00 (0)	0.00 (0)	0.00 (0)
dst<-100	1.00 (63)	0.00 (0)	0.00 (0)	0.00 (0)

nominal t + 12

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	1.00 (49672)	0.00 (178)	0.00 (3)	0.00 (0)
-20>dst>-50	0.91 (2403)	0.09 (228)	0.00 (10)	0.00 (0)
-50>dst>-100	0.79 (296)	0.18 (68)	0.03 (13)	0.00 (0)
dst<-100	0.70 (44)	0.27 (17)	0.03 (2)	0.00 (0)

weight t + 12

	dst>-20	-20>dst>-50	-50>dst>-100	dst<-100
dst>-20	0.92 (45778)	0.07 (3575)	0.01 (386)	0.00 (114)
-20>dst>-50	0.62 (1637)	0.29 (765)	0.08 (204)	0.01 (35)
-50>dst>-100	0.47 (178)	0.29 (108)	0.22 (82)	0.02 (9)
dst<-100	0.29 (18)	0.33 (21)	0.33 (21)	0.05 (3)

TIFPA

deepPP

Thank you