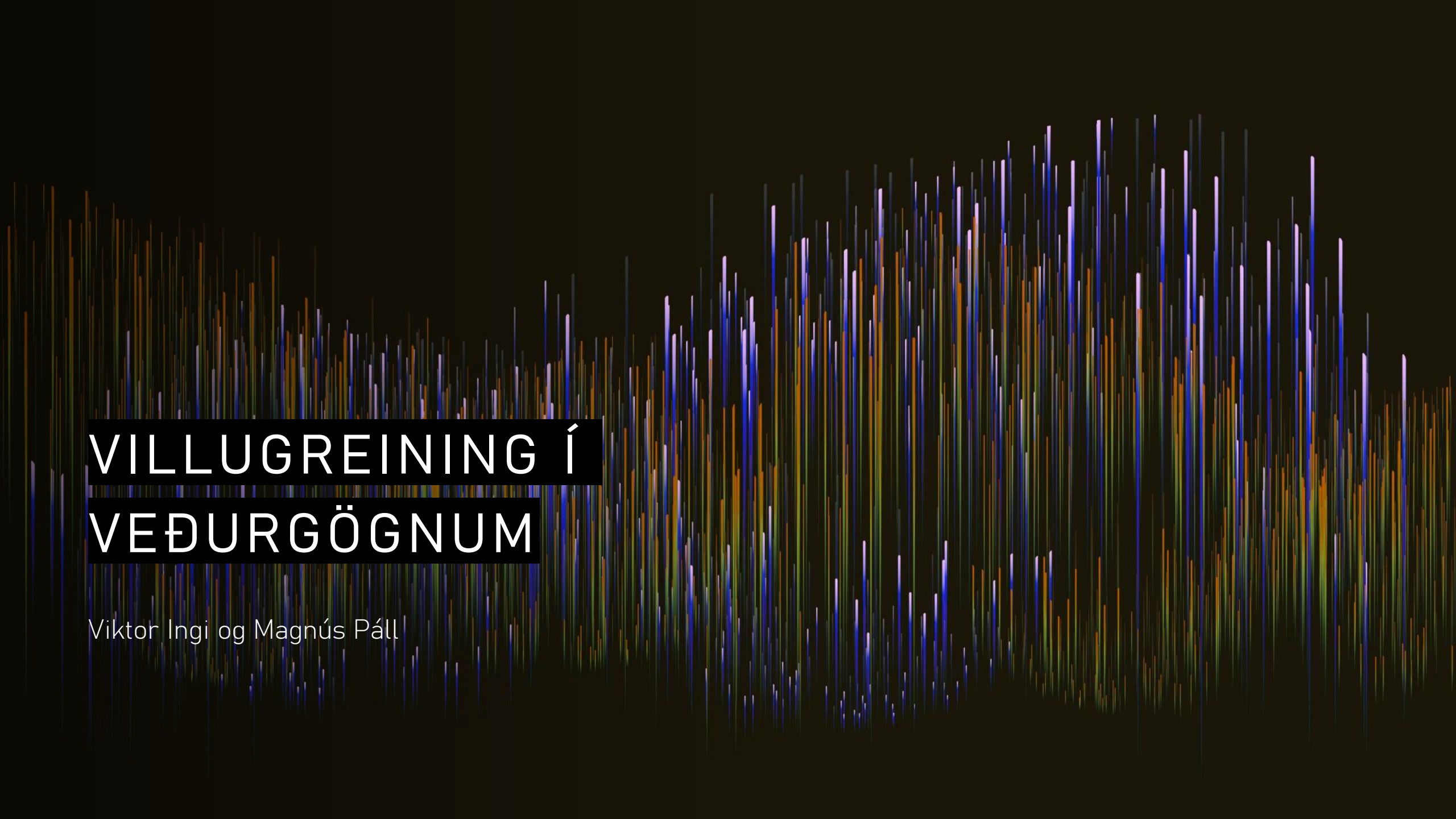
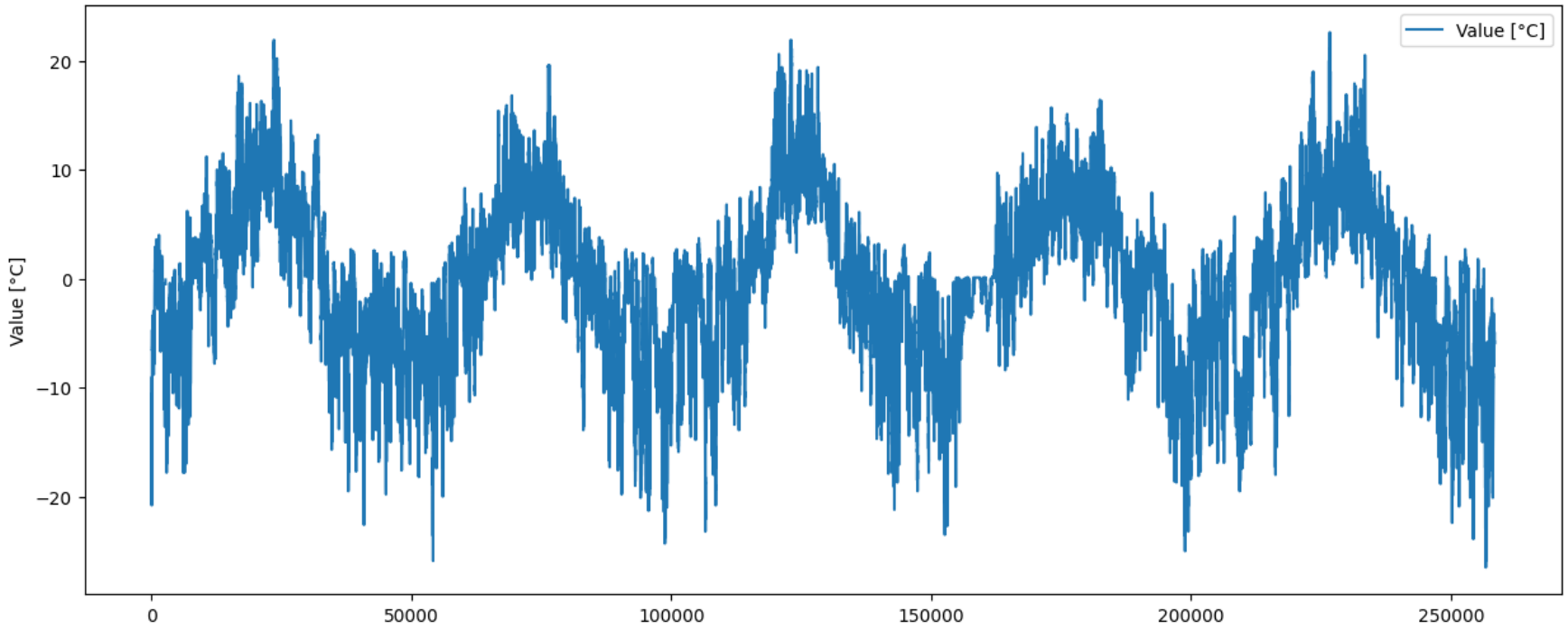


VILLUGREINING Í VEÐURGÖGNUM

Viktor Ingi og Magnús Páll

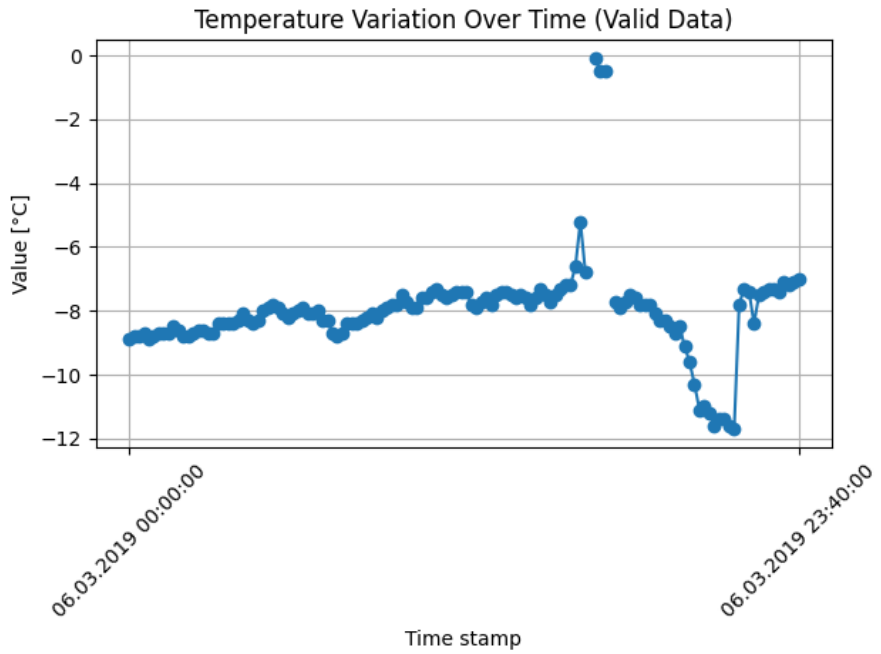
UPPRUNI VERKEFNIS OG TILGANGUR

- Upp á hálendi eru margar veðurstöðvar í eigu Landsvirkjun.
- Í hverri veðurstöð eru mælar af bestu gerð.
- En getur alltaf komið upp svokölluð „spikes“ eða frávik.
- Getur komið upp vegna viðhalds eða villu í mælingu
- Tveir mælar á hverjum stað
  - RTD Pt100 hitamælir
  - RH raka og hitamælir
- Landsvirkjun vilja nota vélrænt nám til að leiðrétta frávik og "spikes" en varðveita samt eins mikið af upprunalegu gildum og hægt er.



MÆLINGAR Í PT100 HITAMÆLIR VIÐ SETUR

- 5 ár af hitastigmælingum
- Febrúar 2019 til febrúar 2024
- Mæling á sér stað á 10 mín fresti og er um 260.000 stök.



	Time stamp	Value [°C]	State of value
2680	06.03.2019 14:40:00	-7.2	200 (Unchecked)
2681	06.03.2019 14:50:00	-7.2	200 (Unchecked)
2682	06.03.2019 15:00:00	-6.6	200 (Unchecked)
2683	06.03.2019 15:10:00	-5.2	200 (Unchecked)
2684	06.03.2019 15:20:00	-6.8	200 (Unchecked)
2685	06.03.2019 16:30:00	NaN	missing (M)
2686	06.03.2019 16:40:00	-0.1	200 (Unchecked), VF
2687	06.03.2019 16:50:00	-0.5	200 (Unchecked)
2688	06.03.2019 17:00:00	-0.5	200 (Unchecked)
2689	06.03.2019 17:20:00	NaN	missing (M)
2690	06.03.2019 17:30:00	-7.7	200 (Unchecked), VF
2691	06.03.2019 17:40:00	-7.9	200 (Unchecked)
2692	06.03.2019 17:50:00	-7.7	200 (Unchecked)
2693	06.03.2019 18:00:00	-7.5	200 (Unchecked)
2694	06.03.2019 18:10:00	-7.6	200 (Unchecked)
2695	06.03.2019 18:20:00	-7.8	200 (Unchecked)

WORST CASE SCENARIO

- Í gagnasetti er bæði missing mælingar ásamt outliers.
- Hér má sjá dæmi um hóp af outliers sem koma vegna viðhalds á veðurstöð.

HREYFANLEGT MEÐALTAL

- Gluggastærð er 9
- Reiknað meðaltal
- Reiknað staðalfrávik
- ± 4 staðafrávik frá meðaltali er merkt sem spike.

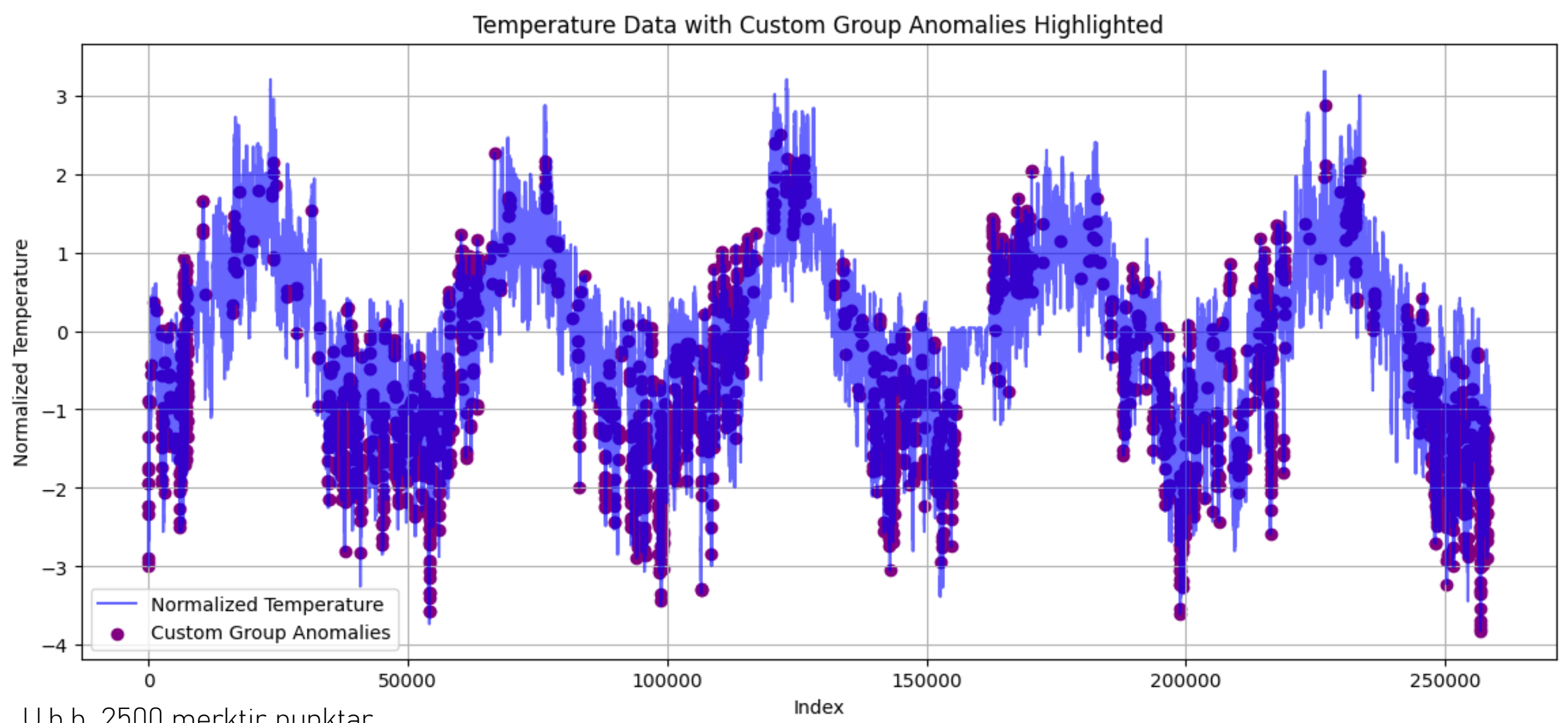
Month	Temp. (°F)	Moving average
Jan	39	
Feb	42	
Mar	50	44
Apr	60	51
May	71	60
Jun	79	70
Jul	85	78
Aug	81	82
Sep	76	81

```
# Reikna moving median og frávik til að skynja outliers.
window_size = 9
df['Moving Median'] = df['Normalized T (degC)'].rolling(window=window_size, min_periods=1).median()
df['Deviation from Median'] = np.abs(df['Normalized T (degC)'] - df['Moving Median'])

# Tökum meðaltal og staðalfrávik af fráviki af median
mean_deviation = df['Deviation from Median'].mean()
std_deviation = df['Deviation from Median'].std()

# Set threshold sem meðaltal plús 4 sinnum staðalfrávik.
deviation_threshold = mean_deviation + 4 * std_deviation

df['Custom Group Anomaly'] = df['Deviation from Median'] > deviation_threshold
```

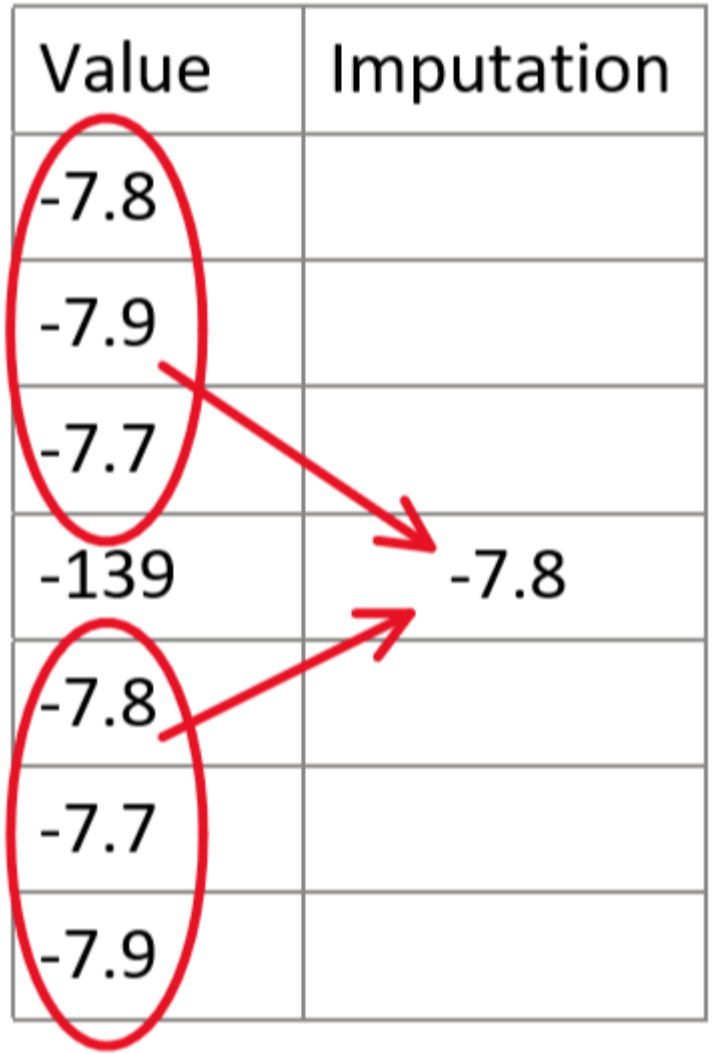


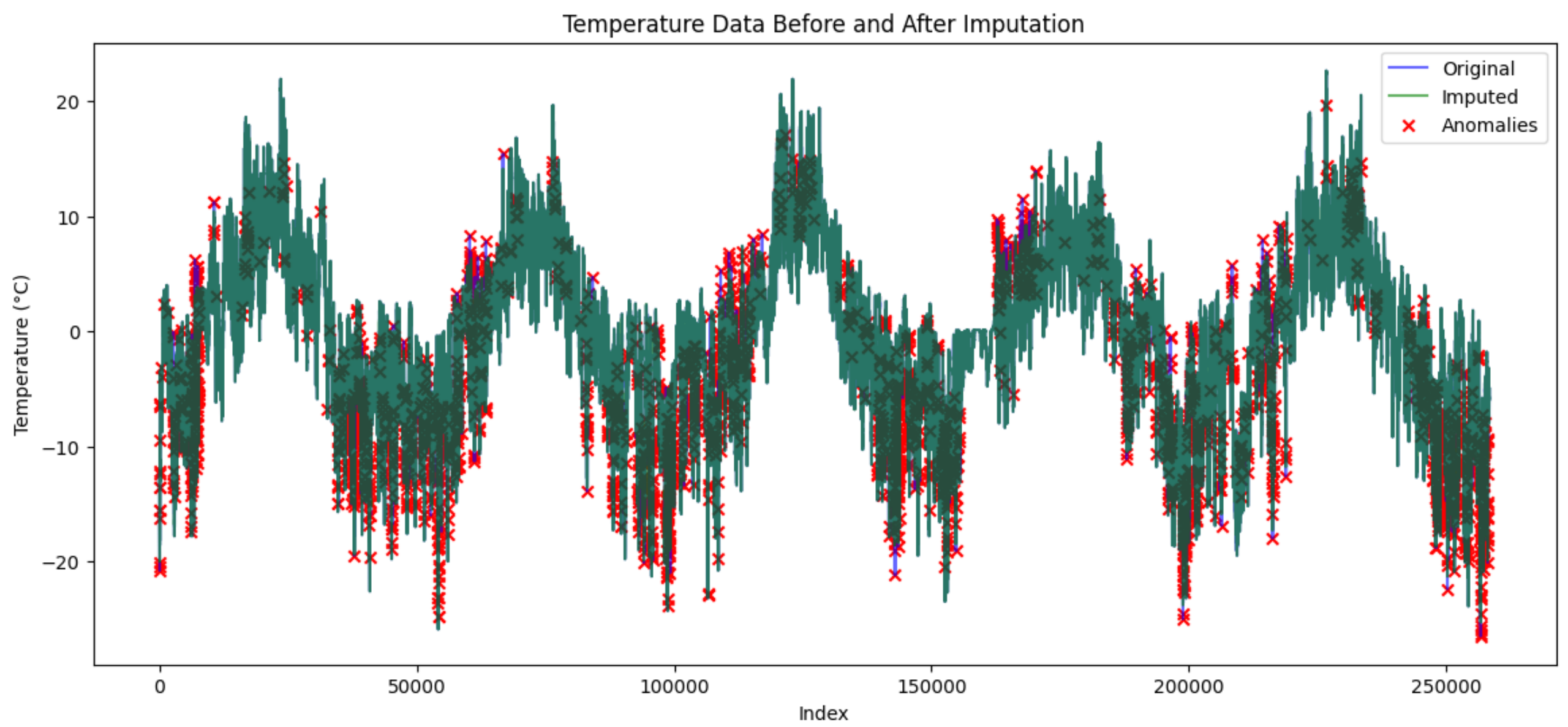
U.þ.b. 2500 merktir punktar

PT100 gögn með merktum útlægum toppum

IMPUTATION

- Laga útlæga toppa
- Reiknum meðaltal næstu nágranna
- Nágrannar eru 5 í hvora átt
- Ef nágranni er einnig útlægur toppur þá er hann hunsaður
- Skiptum svo út útlægum toppum fyrir reiknað meðaltal



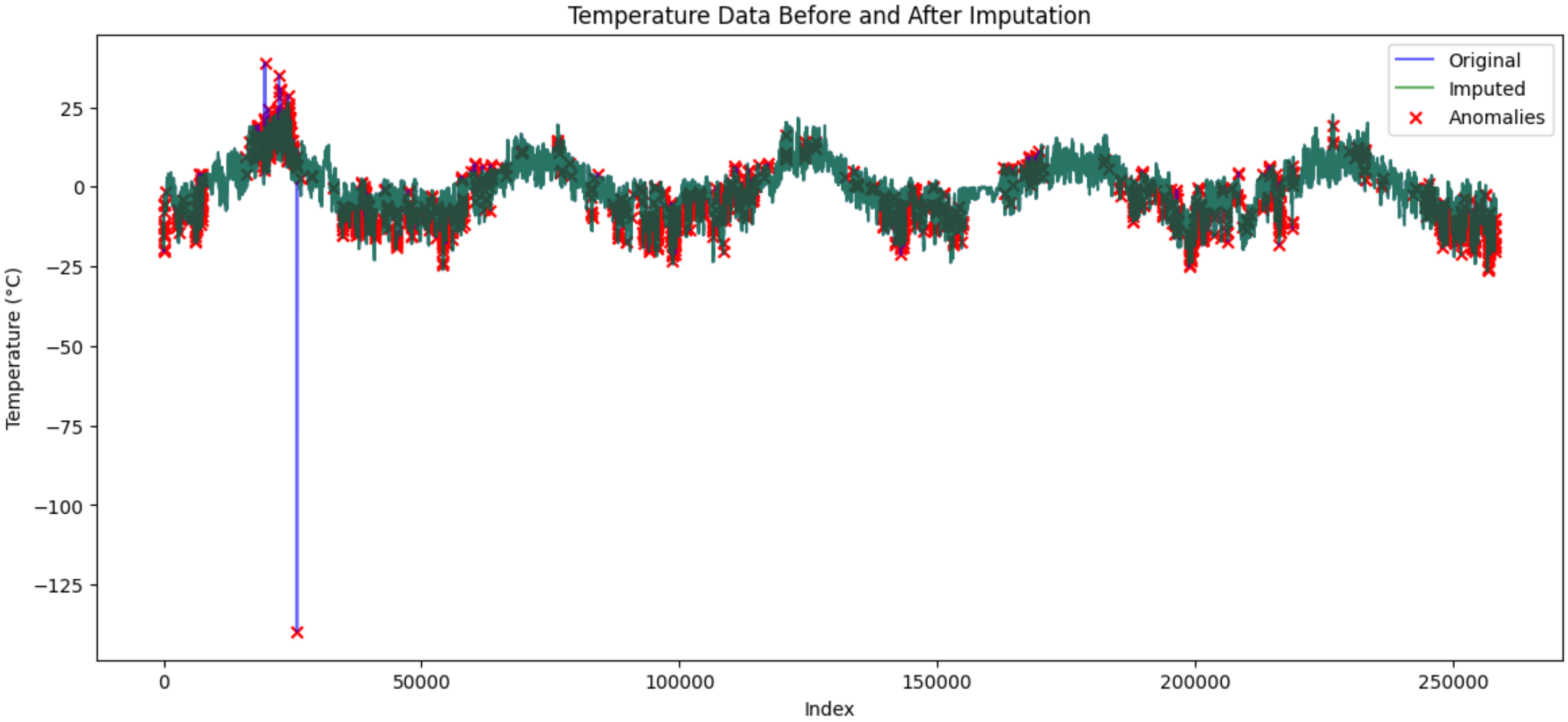


Samaburður á PT100 gögnum fyrir og eftir imputation

Worst case scenario eftir imputation

	Time stamp	Value [°C]
2680	06.03.2019 14:40:00	-7.2
2681	06.03.2019 14:50:00	-7.2
2682	06.03.2019 15:00:00	-6.6
2683	06.03.2019 15:10:00	-5.2
2684	06.03.2019 15:20:00	-6.8
2685	06.03.2019 16:40:00	-0.1
2686	06.03.2019 16:50:00	-0.5
2687	06.03.2019 17:00:00	-0.5
2688	06.03.2019 17:30:00	-7.7
2689	06.03.2019 17:40:00	-7.9
2690	06.03.2019 17:50:00	-7.7
2691	06.03.2019 18:00:00	-7.5
2692	06.03.2019 18:10:00	-7.6
2693	06.03.2019 18:20:00	-7.8
2694	06.03.2019 18:30:00	-7.8
2695	06.03.2019 18:40:00	-7.8

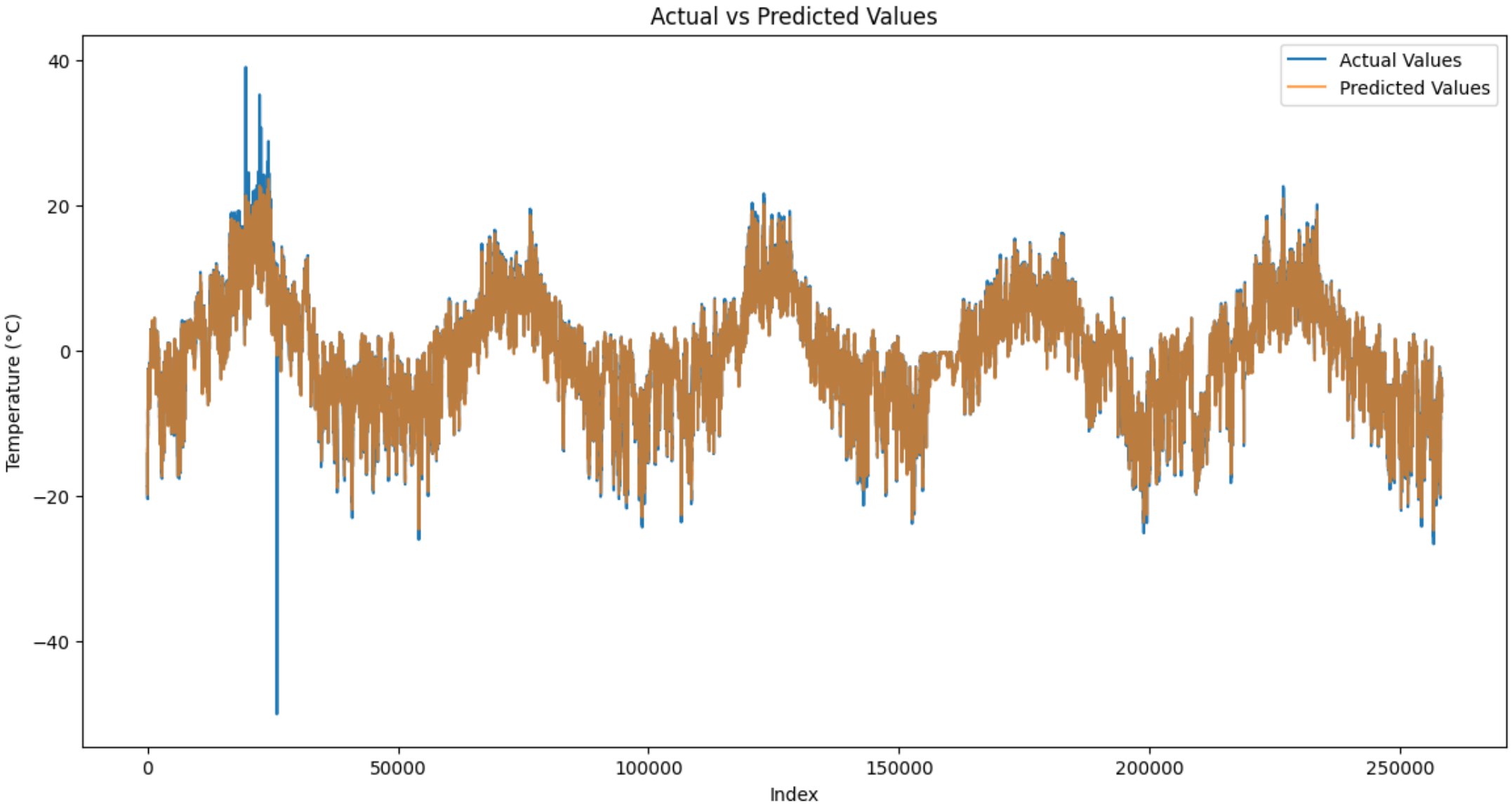
	Custom Group Anomaly	Imputed Value [°C]
2680	False	-7.200000
2681	False	-7.200000
2682	False	-6.600000
2683	False	-5.200000
2684	False	-6.800000
2685	True	-6.900000
2686	True	-6.983333
2687	True	-7.133333
2688	False	-7.700000
2689	False	-7.900000
2690	False	-7.700000
2691	False	-7.500000
2692	False	-7.600000
2693	False	-7.800000
2694	False	-7.800000
2695	False	-7.800000



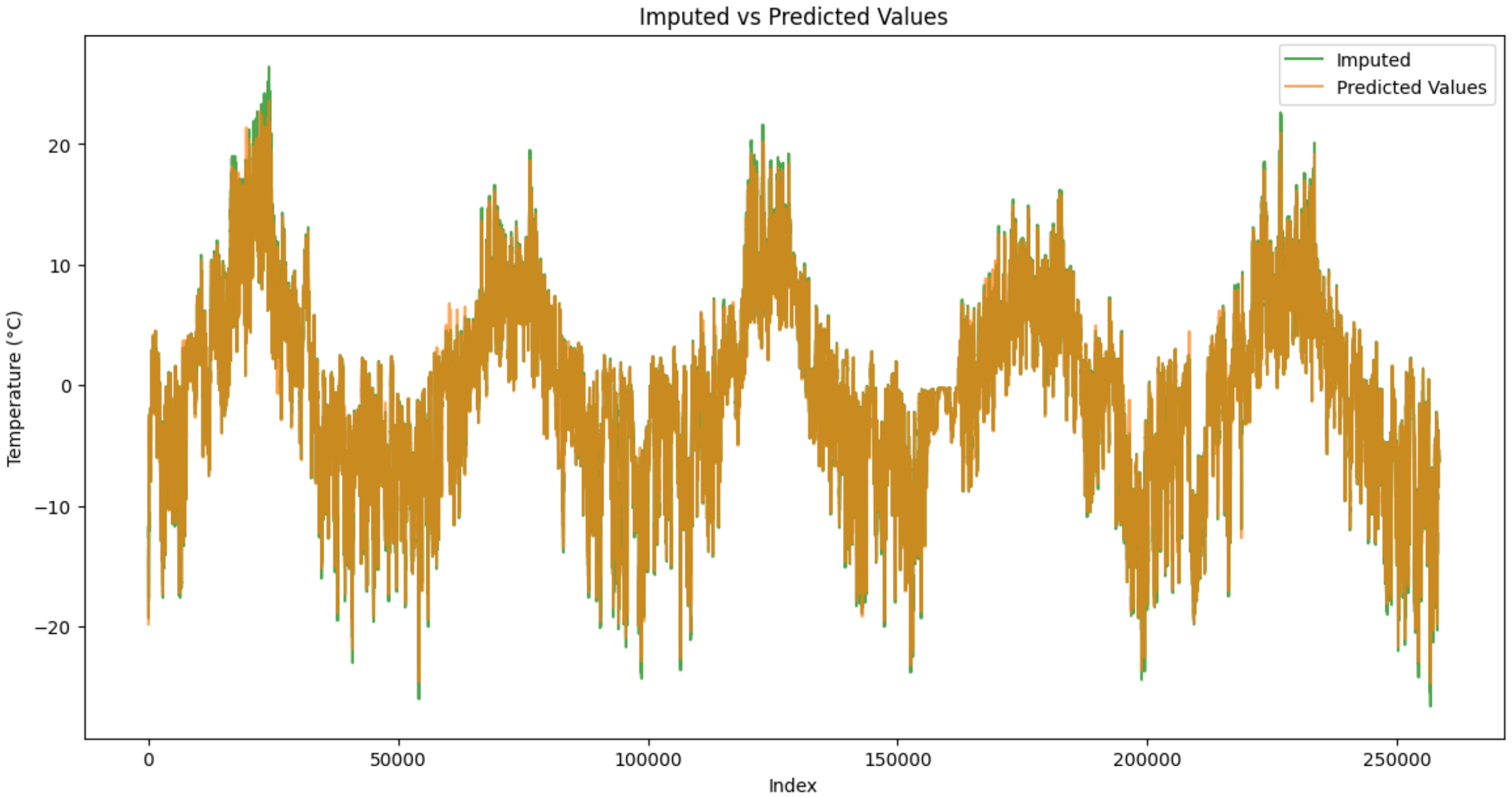
RH mæligögn eftir imputation

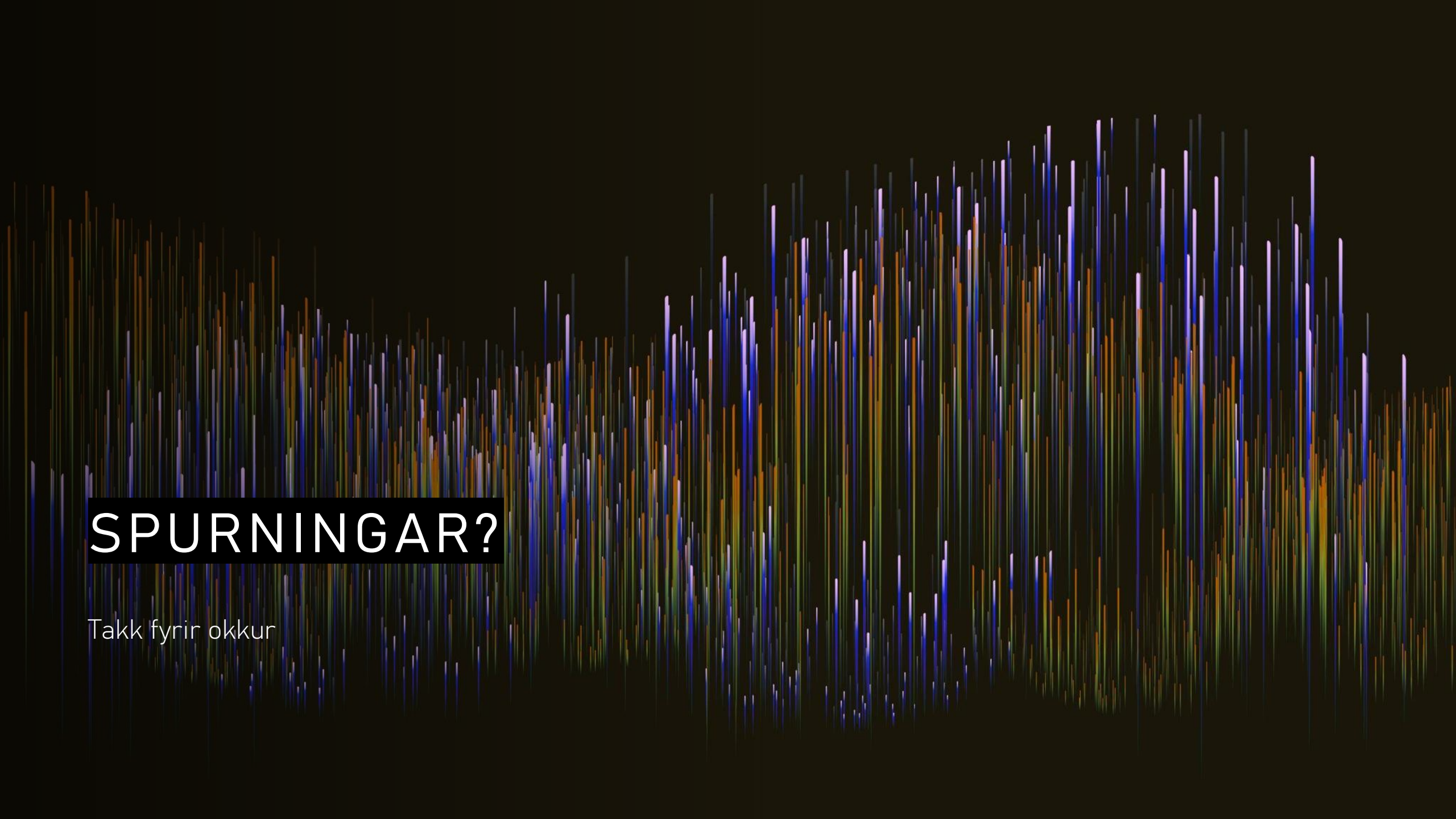
U.þ.b. 1700 merktir punktar

RECURRENT NEURAL NETWORK



Samanburður á RNN og imputation á RH gögnum




SPURNINGAR?
Takk fyrir okkur