A Study of the Dependence of Electron Trigger Efficiency Corrections on the Electron ID

By: Chloe Hooper

(Royal Holloway University of London)

Supervised By:

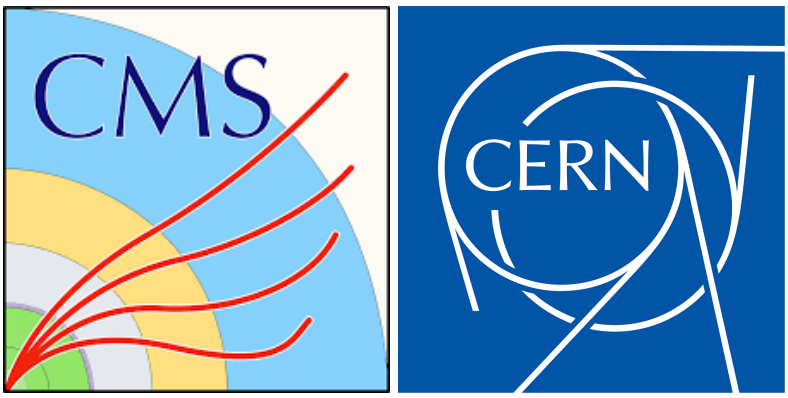
Riccardo Salvatico (riccardo.salvatico@cern.ch)

and Debabrata Bhowmik (debabrata.bhowmik@cern.ch)

CERN Summer Student Program 2024

Abstract

Investigating the importance different ID choices on scale factor on the efficiency of a specific trigger with a focus on the studies from the E/Gamma group. By utilising the tag and probe framework the final outcome of this project was to investigate the impact of choices of IDs on the trigger SFs and to examine the possiblity of providing a universal ID.



Contents

1 Introduction

Within the CMS collaboration the Egamma group work to develop a large array of scale factors (SFs) which can be then applied to the simulated data to improve its agreement with the real collision data. There are many different versions of the scale factors, which aim to correct the fluctuations in how the particles have been reconstructed and identification. These scale factors contain their own errors within them and the purpose of this investigation is to see if any of these SFs have differences which can be accounted for by the uncertainties on these values. Furthermore from this work; is it then possible to draw conclusions on the usefulness of creating so many scale factors and can the number be cut down.

2 Tag and Probe and Scale Factors

The tnp framework and scale factors are a vital part of this report, as such it is important to specify that the high level trigger being used is:

HLT.Ele30.WPTight.GSF There are 5 key IDs that are being compared over the course of this report, those are:

Tight : Cut-Based Tight ID

Medium : Cut-Based Medium ID

Loose : Cut-Based Loose ID

ISO90 : MVA-Based ID 90% efficiency

ISO80 : MVA-Based ID 80% efficiency

These IDs are the efficiencies that we wish to probe against. The cut based IDs are specified via shower, shape and isolation conditions which are placed over multiple variables chosen when the IDs are made. There are also the MVA based IDs which utilises a multi-variable algorithm to determine the selection criteria.

2.1 Tag and Probe

To begin, it is important to go over the basics of the tnp framework that is being used for the SFs comparisons. The tnp framework utilises the well-studied mass resonance of the Z boson to probe an efficiency; the goal of this is to use only true electrons in data and ignore any fakes. [1] A tag is assigned to an electron that has

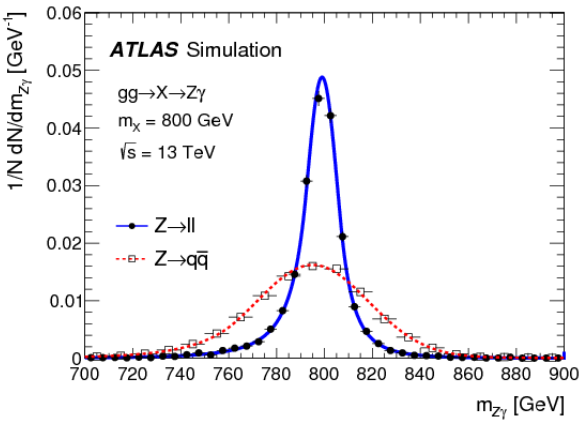


Figure 2: Z Boson mass resonance graph

passed a certain criteria, another electron is then selected to be used as the probe. These probes are used to verify the efficiencies being analysed. This is then repeated many times until there is a final value that can be assigned to this method. For the purpose of this investigation the formula for efficiency is defined as follows:

$$\frac{passing\ probes}{passing\ +\ failing\ probes} \tag{1}$$

This method is applied over a certain trigger to find its efficiency that passes further selection criterion these further selections which are referred to as ID that has been applied. There is also more specification in the

values that the tags can come from in the data such as "tags $|\eta|$ cut=2.17" and r9=0.98 both of which will be discussed later in this report.

2.2 Scale Factors

Once that an efficiency has been determined using tnp, it is then important to check that there are no variations in the data due to differences with the monte carlo simulation that may be misleading in results. To do this a scale factor is used. All of the selection criteria are then copied and the tnp framework is then run on a Monte-Carlo simulated data. This Monte-Carlo efficiency is then divided by the real data version to determine its scale factor.

$$\frac{Real\ Efficiency}{Simulated\ Efficiency} \tag{2}$$

This can then be applied to the simulation when it is compared to collision data to assure it is an accurate distribution. These SFs are also dependent on the conditions that are applied to them, this includes the IDs that are referred to earlier.

3 Method

The method that is best to use for this investigation was discussed extensively it was in the end settled on that the comparison should be made from the standard deviations from the error. To do this one bin was subtracted from the other and then divide this value of the average error. The purpose of this is to determine how far away the SFs are from each other in terms of the standard deviation. The final formula for the standard deviation graph is:

$$\frac{Bin\ of\ Scale\ Factor\ with\ ID\ X\ applied - Bin\ of\ Scale\ Factor\ with\ ID\ Y\ applied}{Average\ Uncertainty\ for\ that\ bin\ in\ X\ and\ Y} \tag{3}$$

3.1 Preliminary results

The preliminary study for this investigation used n-tuples that were pre-made and had already had the tnp criterion applied. The files can be found here:
`/eos/cms/store/group/phys_egamma/ScaleFactors/Data2022/ForRe-recoE+PromptFG/tnpEleHLT`
They where promising results with a good spread of bins that had a standard deviation of less than one and only no one area where this was performing worse.
In this version the bins from two IDs being compared were subtracted from each other and then the highest error from the two bins was picked and then divided by the subtracted values. In section 3 it is possible to see the deviations, the histograms also needed to be scaled to match the original ROOT files.
Further discussion of these results with the CMS-EGamma group proposed that these differences may have origins that are due to imperfect fitting on the data rather than a differences in the SFs themselves. This then lead to the conclusion that the next steps for this project was to recompute the SFs using a cut and count approach, rather than fitting; make my own n-tuples, SFs and, apply the ID's.

3.2 Remaking the scale factors

To remake the scale factors, another python egamma-tnp package was utilised, to produce the tnp n-tuples [2]. As this package is currently in the beta stage it still need further work. It was required to improve on this framework so that it could be used to develop 2D efficiencies with the n-tuples and then also further developments using the boost histogram package[3] were completed to create the 2D scale factors. The egamma-tnp package was invaluable for this project and comparisons to the original tnp framework package has shown that match up very well.

3.3 Errors

For the comparisons between SFs to work correctly it was imperative that the errors were being calculated correctly. The uncertainty in the individual efficiencies was calculated using the boost histogram interval model. This function calculates the uncertainty using the Clopper-Pearson confidence interval with the assumption of a Binomial parameter, (centred around it being an efficiency) This however could not be used for the scale factor itself as the scale factor is a ratio of efficiencies and not an efficiency itself. For this reason it was necessary to calculate the SF uncertainty in a different manner, this formula had to be derived using partial derivatives following the standard error propagation method.

$$\frac{\sigma_z^2}{z^2} = \frac{\sigma_x^2}{x^2} + \frac{\sigma_y^2}{y^2} \tag{4}$$

Where z = (x/y) and σ denotes the uncertainty on that value. This was then used to calculate the average error between the two SFs.

3.4 Extra Variables on the scale factors

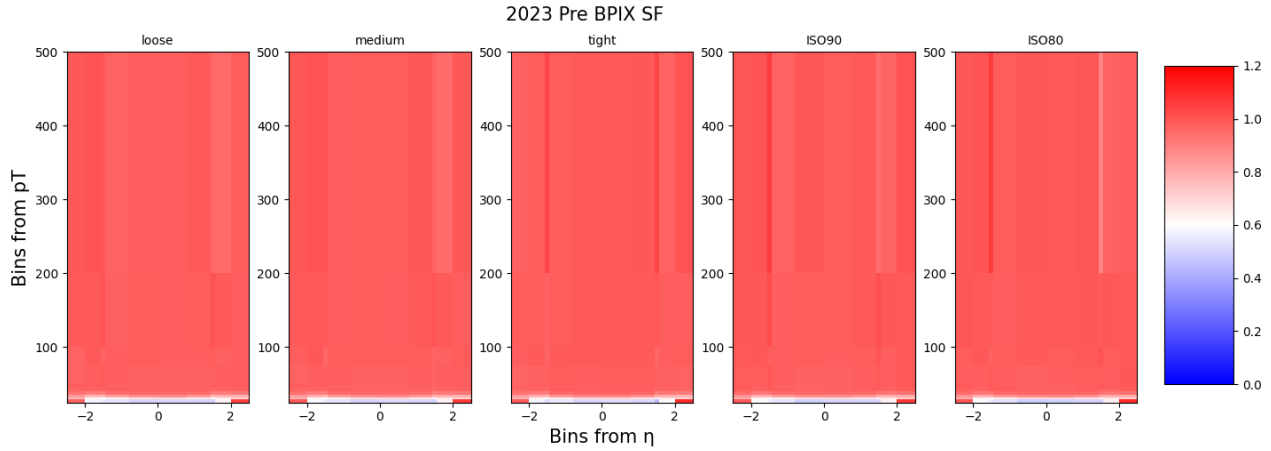
The Scale Factors are not only affected by the already specified IDs but also some extra variables that are r9 $<$ 0.975 , $\eta < 1.44$ or 1.56, dxy < 0.05 or 0.1, dz < 0.1 or 0.2. The eta values have the same meaning as those in the histogram however dxy, dz and, r9 require further explanation. The dxy and dz values are related to the position of the object being passed through the filter with respect to the primary vertex of the pp collision, the position can be used to assume things about the energy of the particle when looking at the bremsstrahlung of the particle , so it is possible to filter by these values if a certain particle is wanted to be isolated by the trigger. The r9 value is related to the energy spread of the particle shower. R9 > 0.975 is requiring that the 97.5% of the energy can be found within a 9 crystal area within the detector around the seed crystal, so its looking for a particle signal with lower bremsstrahlung.

4 Results

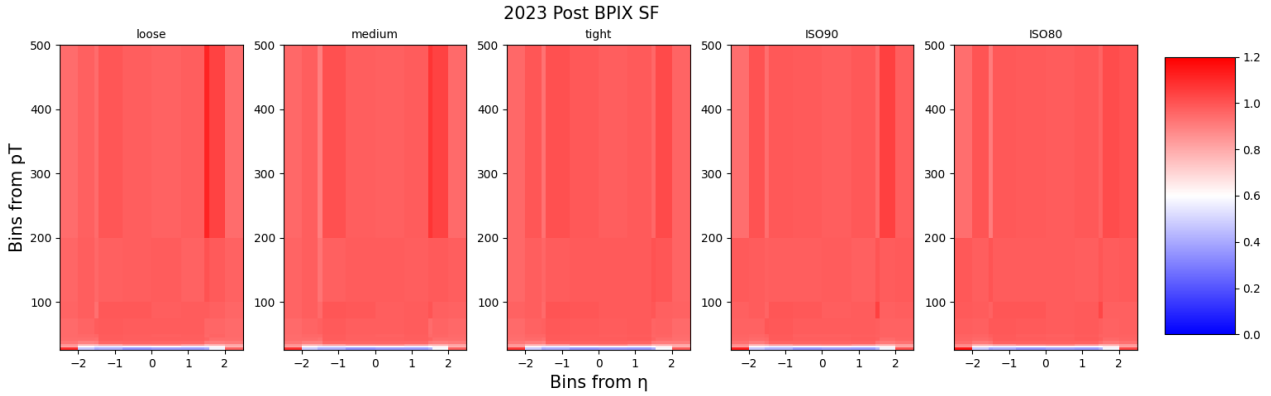
.

4.1 Scale factors

The scale factors are made using the formula above in section 2. The histograms for each file for the 2022 and 2023 data taking periods can be seen in the figures below. The scale factors were remade and then re-plotted using the boost histograms package.



(a) 2022 Pre EE Scale factors



(b) 2022 Post EE Scale factors

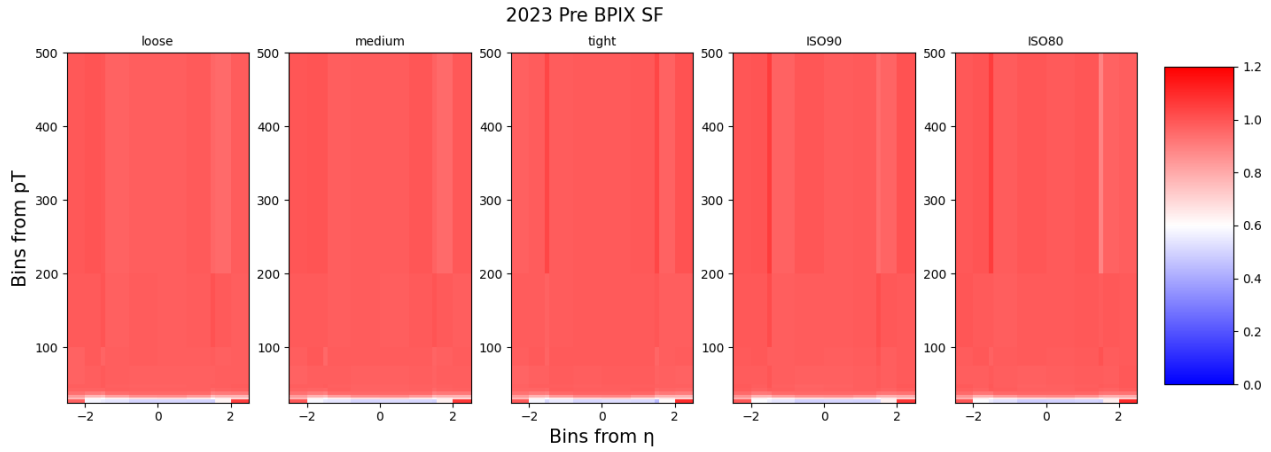
Figure 3: 2022 Scale factors

The data for the year 2022 have been split into the Pre and Post EE eras, this was the time when the ECAL lost function in the ECAL-end cap due to a water leak, leading to two different detector conditions that have to be matched in simulation.

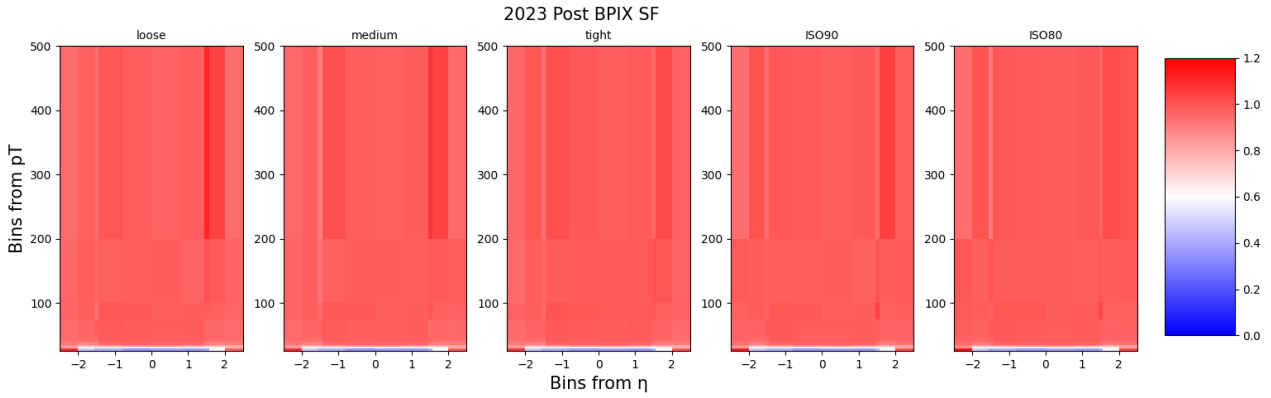
Similarly the data taking for the year 2023 has also been split into two eras. These periods are due issues with the barrel pixel detector, which then necessitated the splitting of the data taking into the two periods.

4.2 Comparisons

The comparison graphs were made by subtracting the the scale factors from one another. The SFs corresponding to the ID on the top along the graphs is the one that is being subtracted from the ID along the left side, these subtractions then occur bin by bin. The differences in between the bins is very low in most cases which will be a factor in the conclusion that will be drawn later in the report.



(a) 2023 Pre BPIX Scale factors



(b) 2023 Post BPIX Scale factors

Figure 4: 2023 Scale factors

A similar case is true for the 2023 data however the differences between this data is even smaller on average than that of the previous years. It is also noticeable that the uncertainty is higher within the barrel to end cap transition region which is expected due to a gap in this detector part.

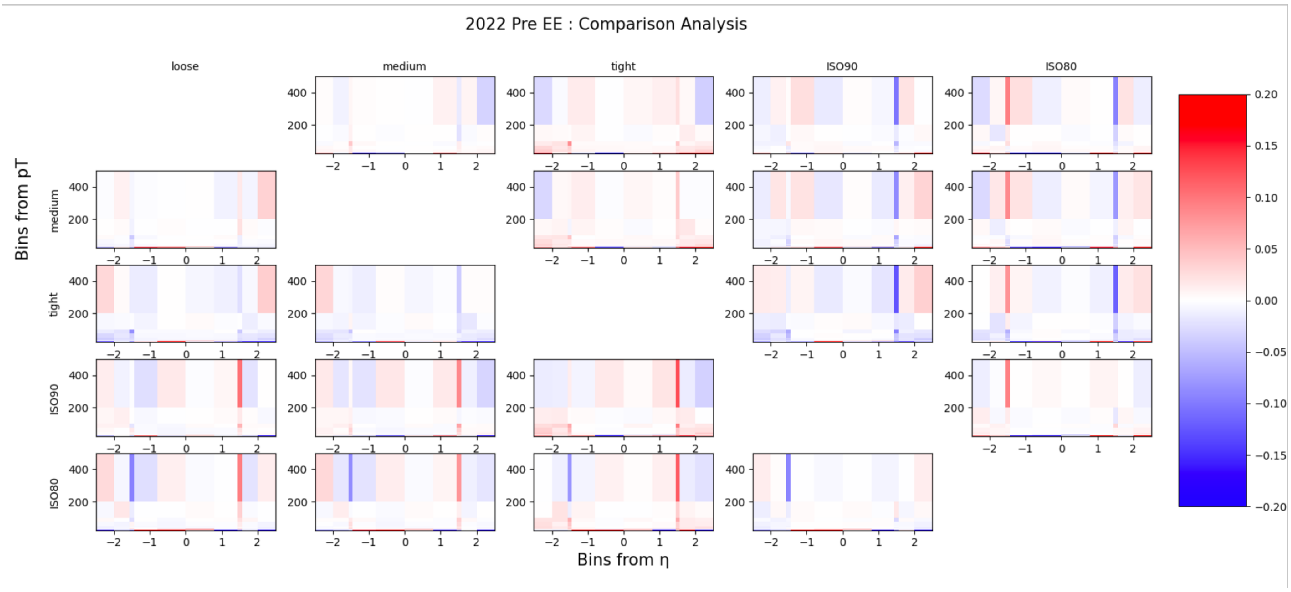
4.3 Errors

The errors are an interesting case between the years, again the areas transitioning to the barrel are highly pronounced within the histograms. The average error was found using the formula discussed above and lead to highly consistent spreads of uncertainties throughout the analysis.

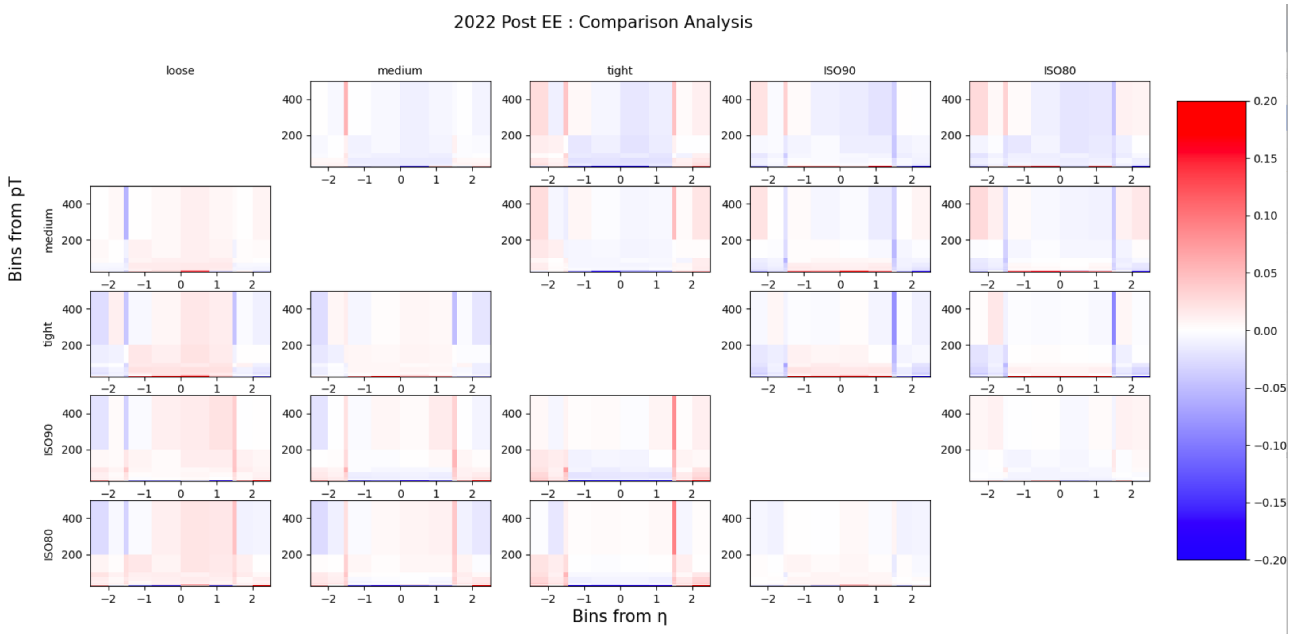
The final step of this analysis is to divide the compared data by these average errors in order to get the final standard deviation histograms. This can then be used to see if the relation between these SFs can be considered to be within the uncertainty between the values.

4.4 Standard Deviation

The standard deviations that can be observed within these histograms is very stark in some places, and continues to be across all the data taking eras.



(a) 2022 Pre EE Comparisons

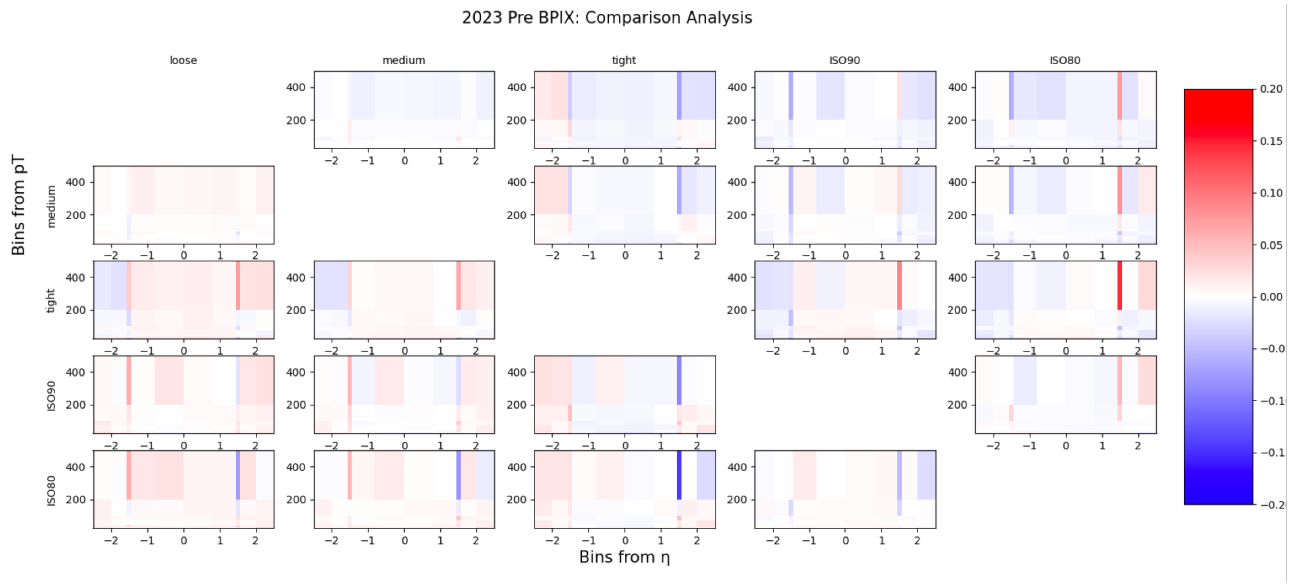


(b) 2022 Post EE Comparisons

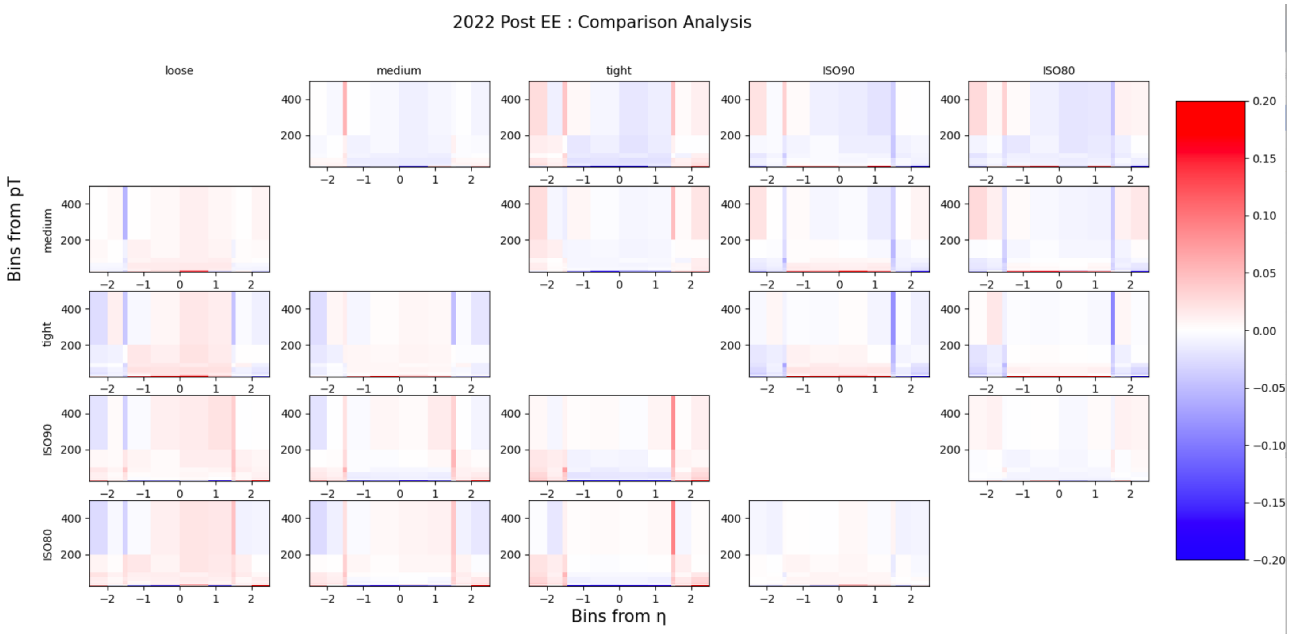
Figure 5: 2022 Comparisons

The first half of 2022 data has less values with a standard deviation above one however it does still contain multiple values with a standard deviation of 3 or above, for all of these graphs the maximum values for standard deviation of reaches above 10 when the analysis focuses on the barrel to end cap transition period in eta and a lower pT area.

The standard deviations of the 2023 data is also indicative of the scale factors being different for different ID



(a) 2023 Pre BPIX Comparisons



(b) 2023 Post BPIX Comparisons

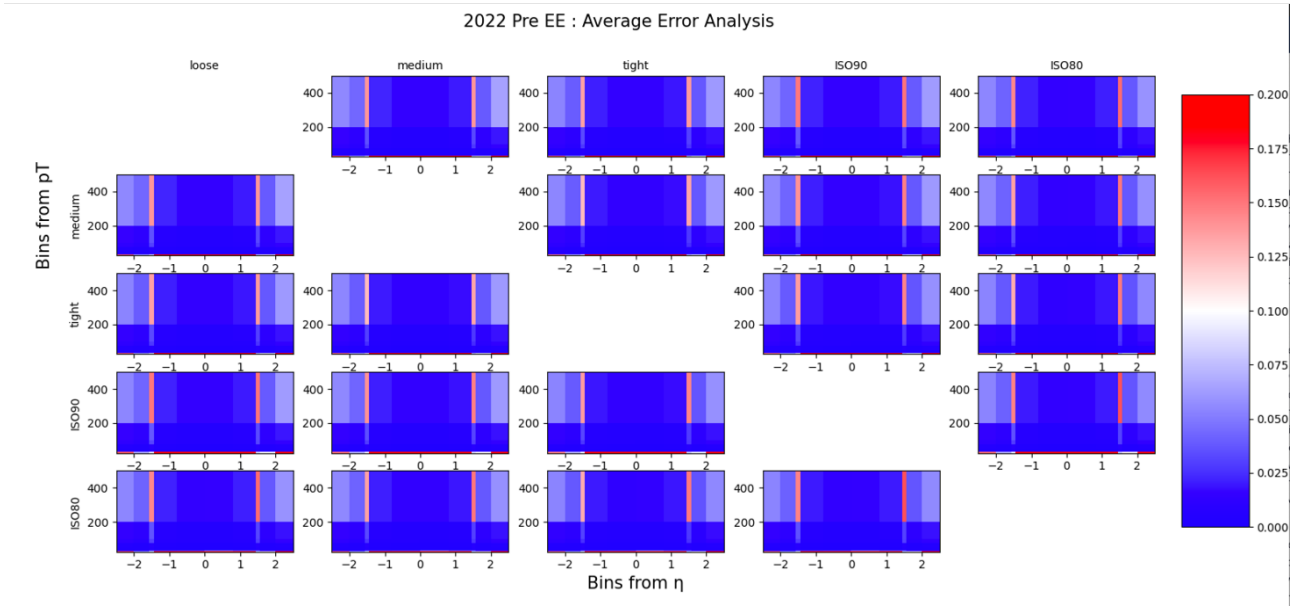
Figure 6: 2023 Comparisons

choices. However it is possible to see that there is significantly less fluctuations across the loose and medium ID types.

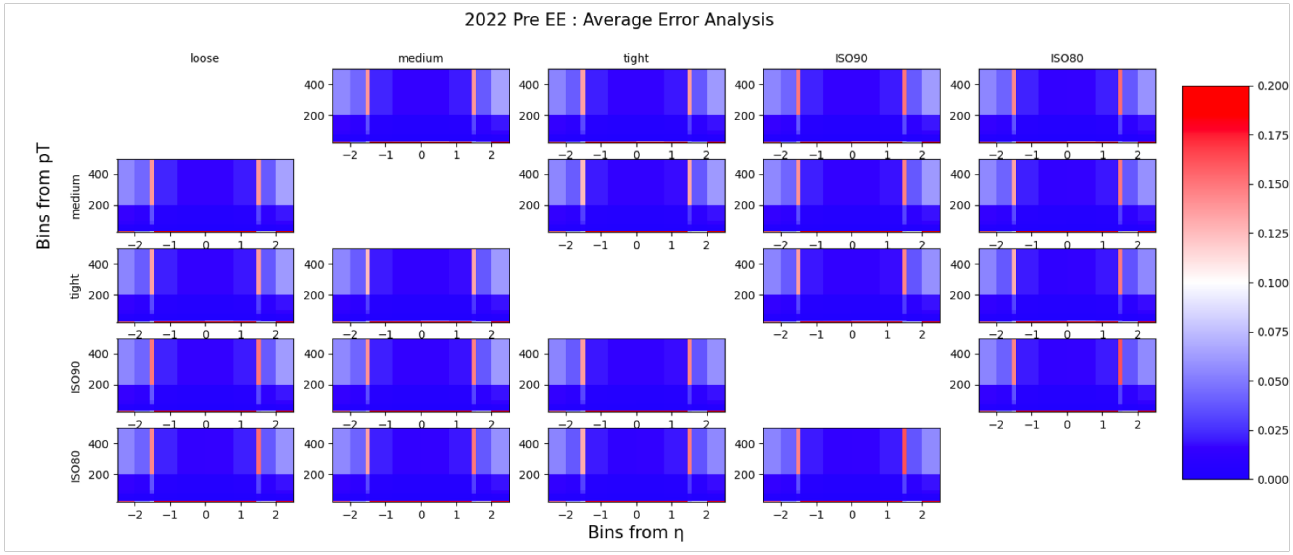
4.5 Extra Parameters and Deeper Analysis

To support the preliminary results it was also important to repeat the process of further fitted data.

These extra conditions were those that were specified above. Specifically the r9 > 0.975 , $\eta < 1.56$, dxy $<$ 0.1, dz < 0.2.



(a) 2022 Pre EE Averaged Errors



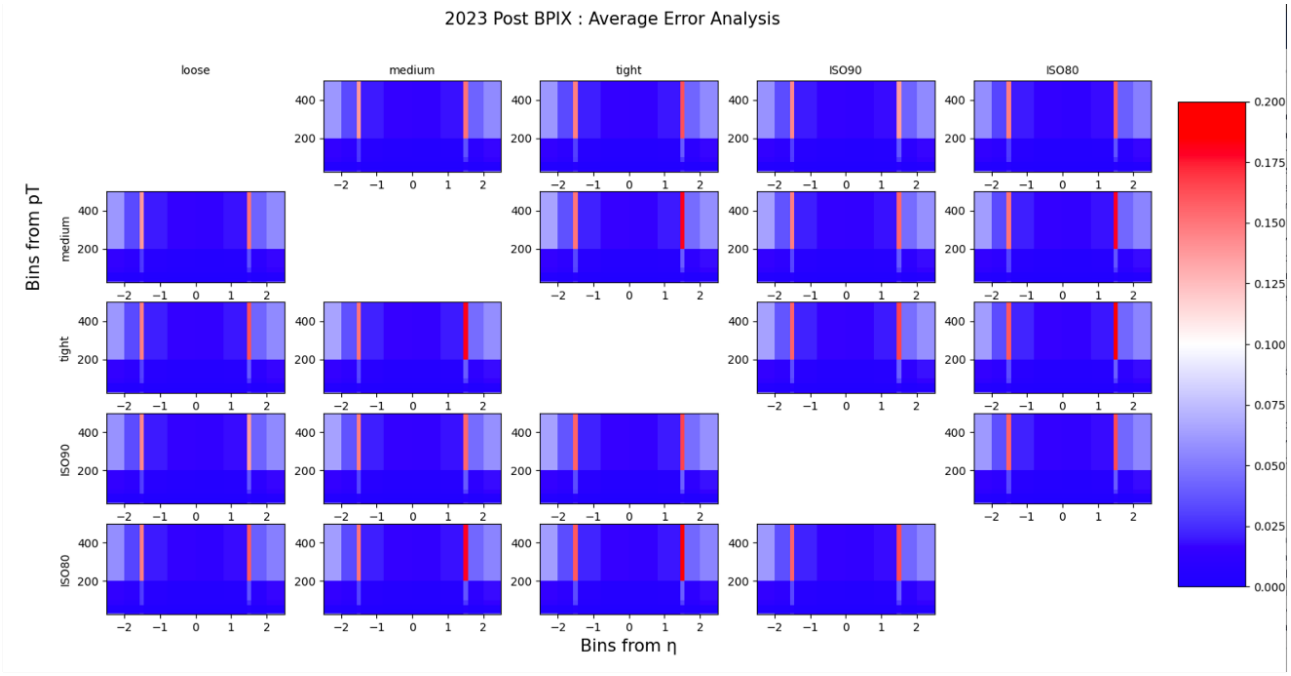
(b) 2022 Post EE Averaged Errors

Figure 7: 2022 Averaged Errors

The final distribution does lead to the same pattern for high standard deviation that was observed in the original results.

The efficiency in the real and simulated data can be seen in the histograms below. Both tnp efficiency graphs look as expected.

(a) 2023 Pre BPIX Averaged Errors

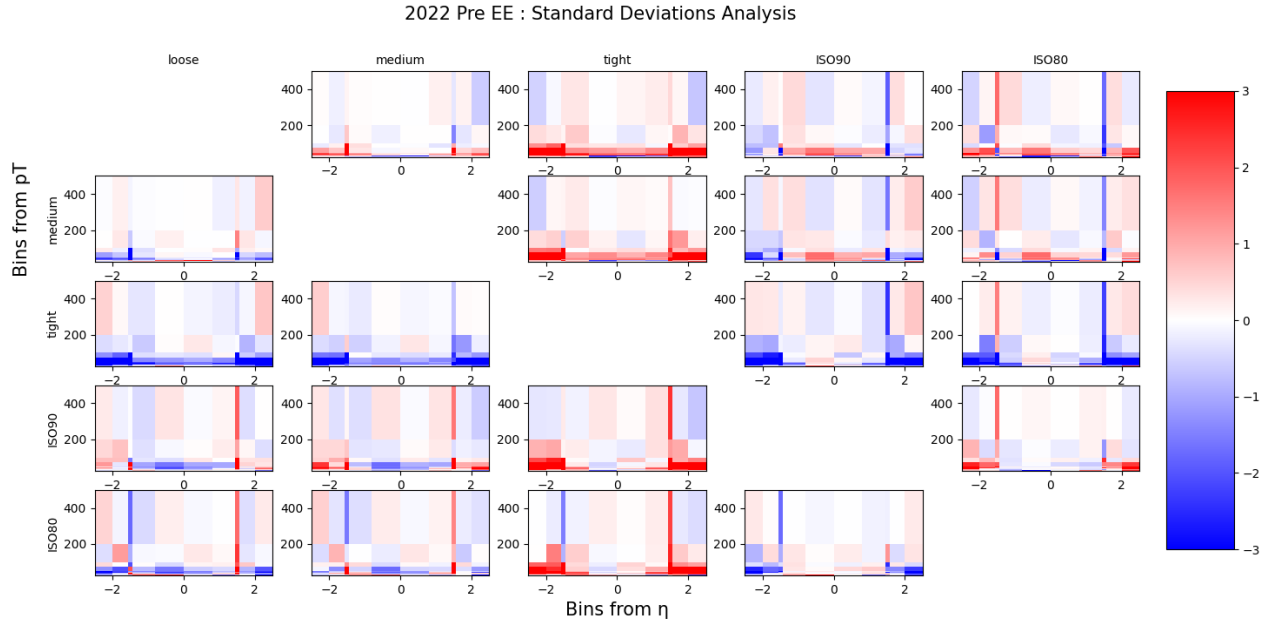


(b) 2023 Post BPIX Averaged Errors

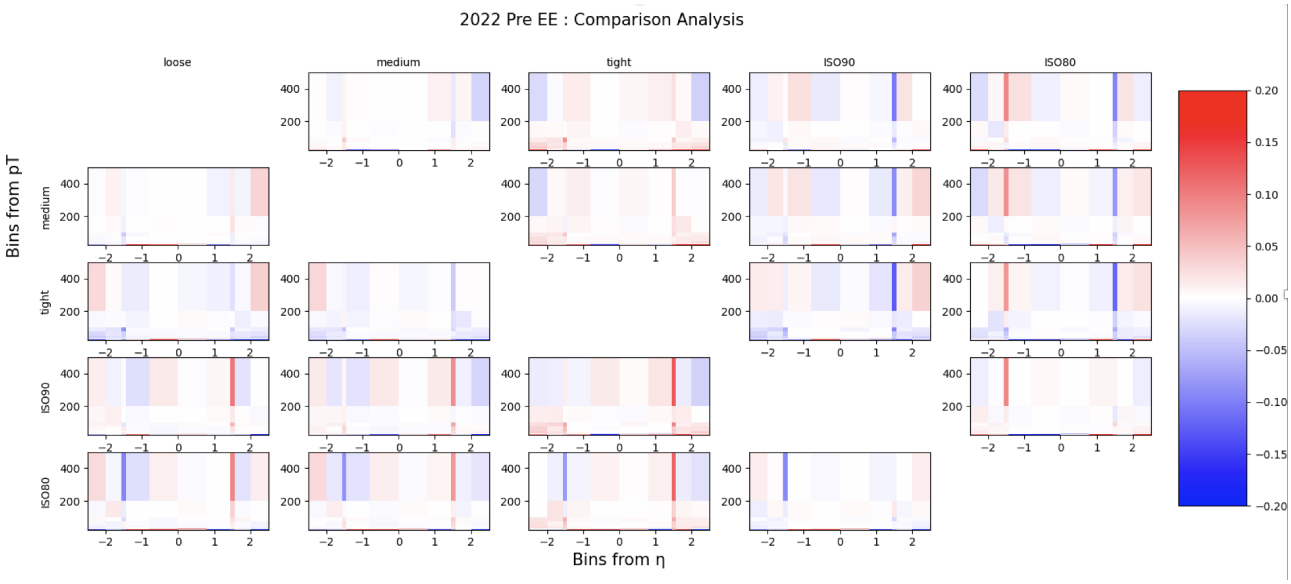
Figure 8: 2023 Averaged Errors

5 Conclusions

The purpose of this study was to determine the impact that the ID type has on the SFs. The comparison between central values divided by the uncertainties. To conclude, the impact of the ID choices on the SFs is non-negligible as can be seen in 2. Their distribution and standard deviation is greater than 2 which indicate



(a) 2022 Pre EE Standard Deviation



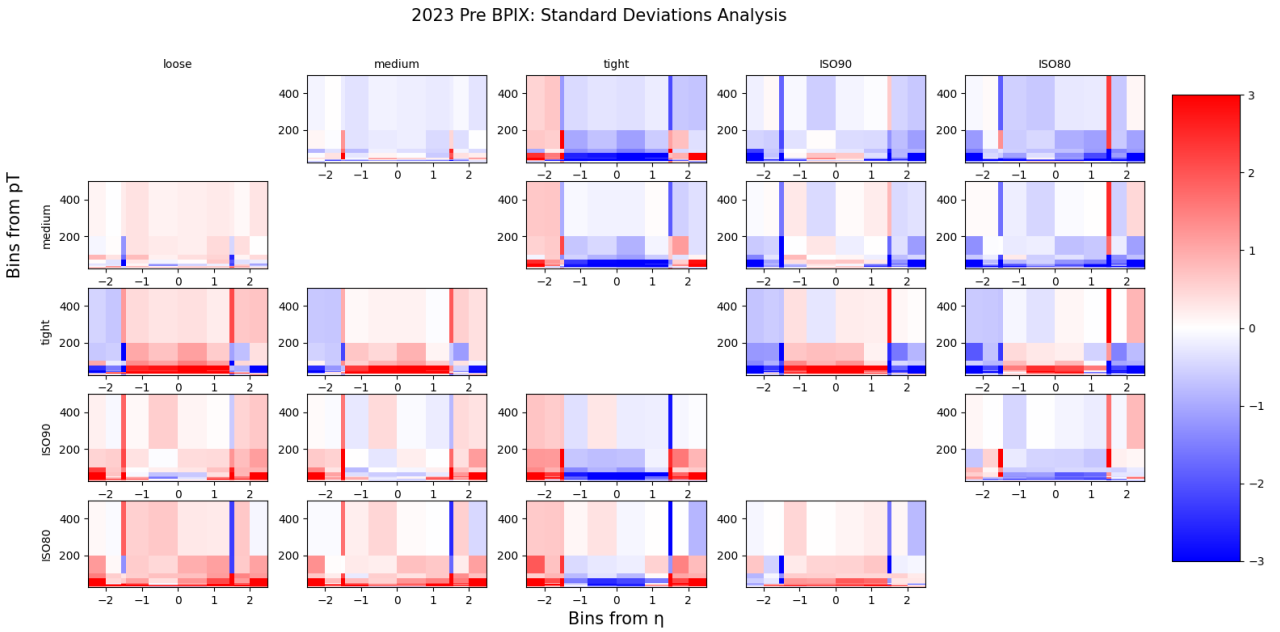
(b) 2022 Pre EE Standard Deviation

Figure 9: 2022 Standard Deviation

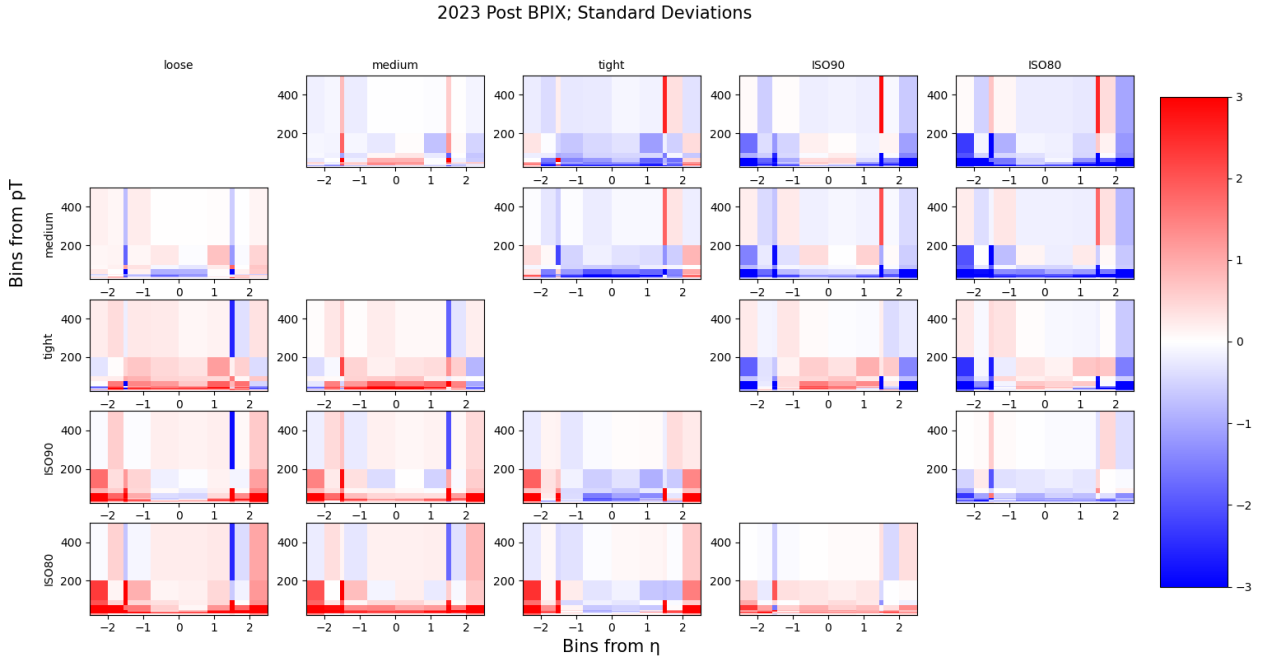
that there is a considerable amount of phase space (pT-η bins) the nominal values of SFs due to two different ID choices are not covered by the uncertainties of the SFs.

6 What has been learnt

There has been significant amounts of knowledge that has been gained over the course of this project including CMS HLT, ID, use of the TnP framework, ROOT, coffea, conda, and managing large datasets.



(a) 2023 Pre BPIX Standard Deviation



(b) 2023 Post BPIX Standard Deviation

Figure 10: 2023 Standard Deviation

7 Acknowledgements

I would like to say a massive thank you to my supervisors Riccardo and Debabrata who have been so supportive over this project. They have been really amazing to work with and an invaluable resource over the course of the summer. I would also like to thank Iason for his patience and guidance with the Egamma-tnp package and, Logan and Nhi for being my desk buddies. Finally; I would like to thank the CERN summer project team for

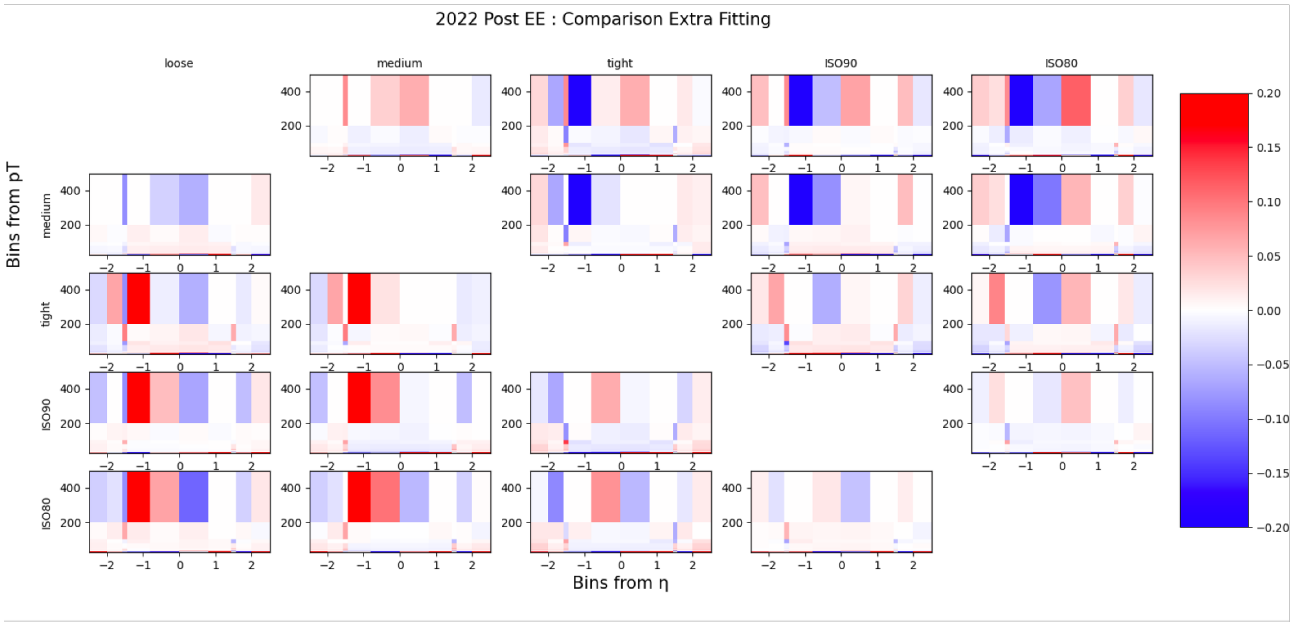


Figure 11: 2022 Pre EE with extra parameters Comparisons

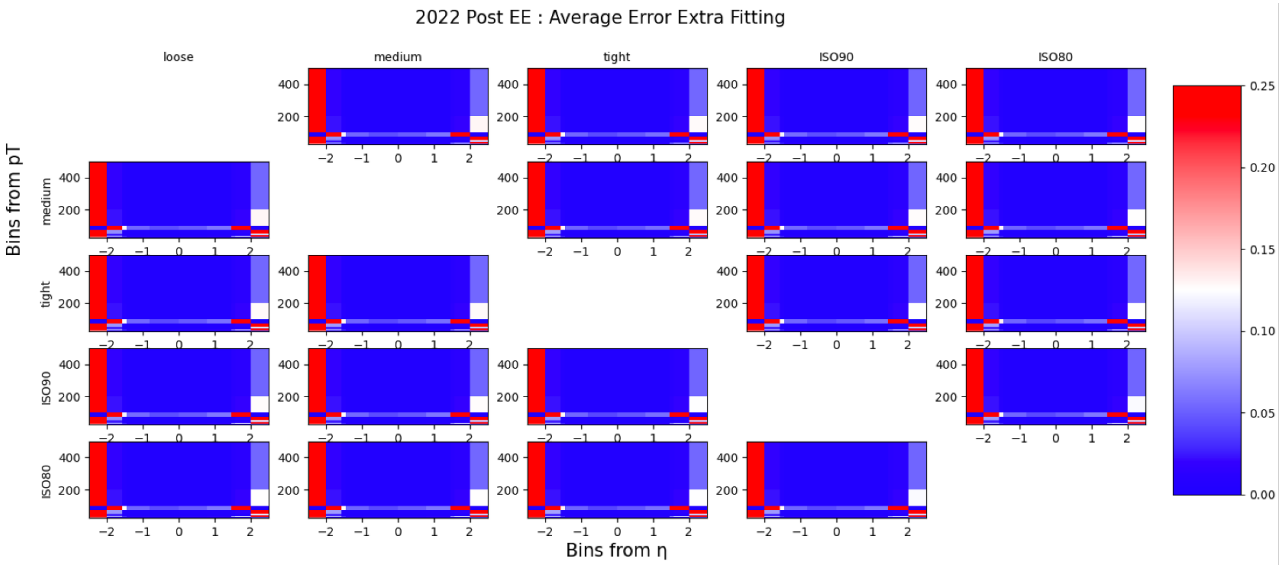
all their work making this wonderful summer possible.

References

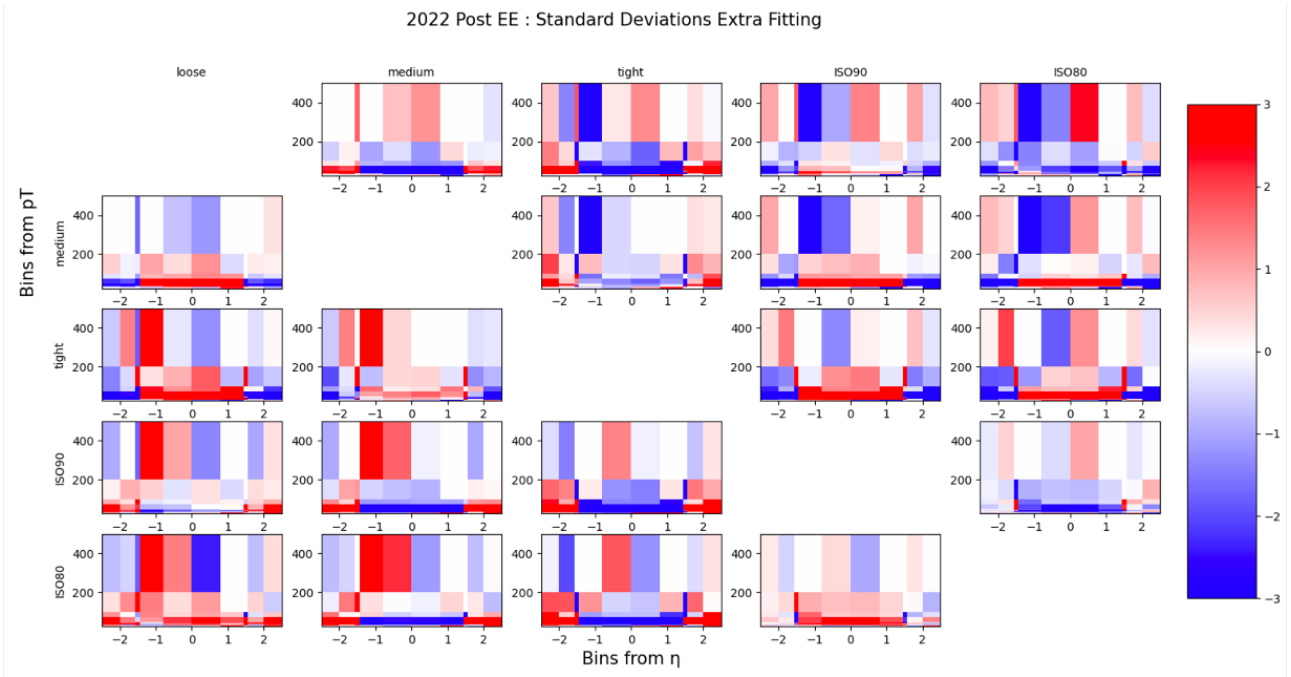
[1] CMS Collaboration, *Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC*, CMS-EGM-17-001, Section 4.7 (2021)

[2] Iason Krommydas, *E/Gamma High Level Trigger efficiency from NanoAOD using Tag and Probe in the coffea framework.*, `https://gitlab.cern.ch/cms-egamma/egamma-tnp/-/tree/master` (2024)

[3] Henry Schreiner and Hans Dembinski *Boost Histogram Package*,`https://boost-histogram.readthedocs.io/en/latest/api/boost_histogram.html` (2024)

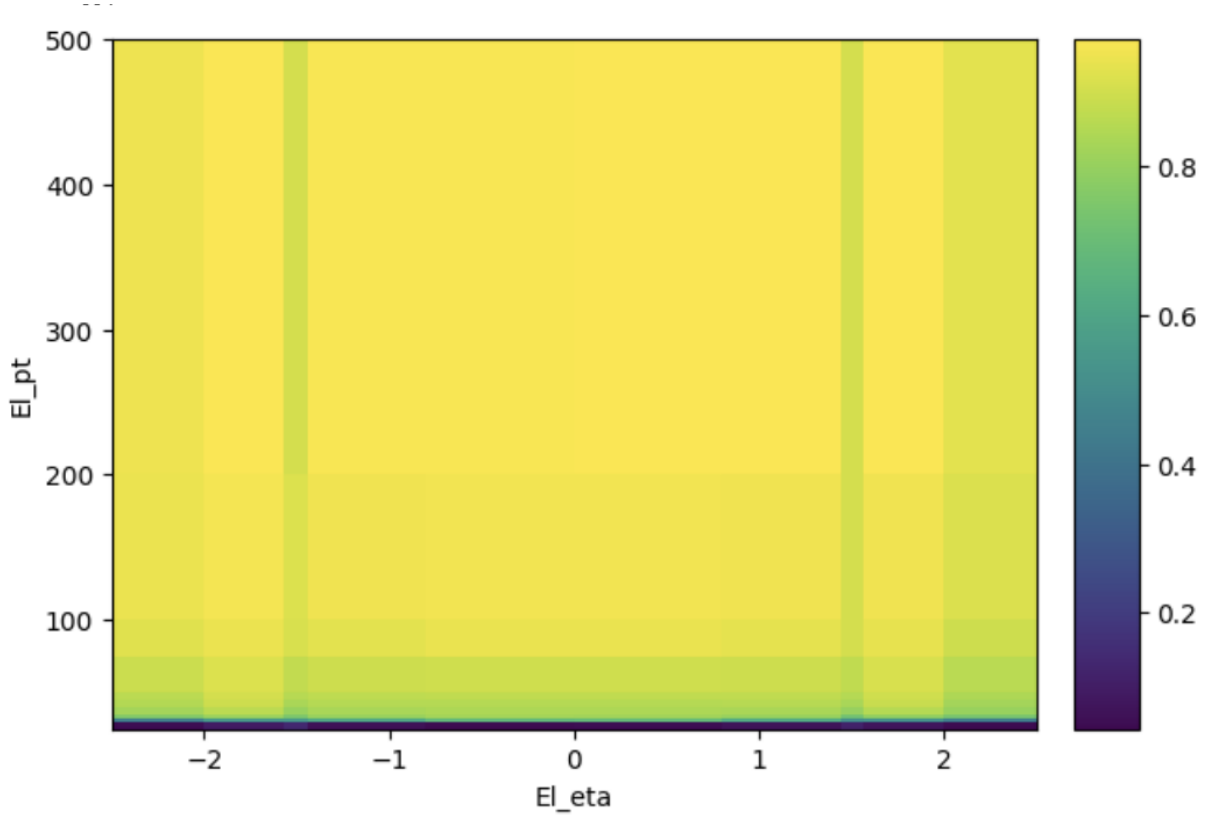


(a) 2022 Pre EE with extra parameters average errors

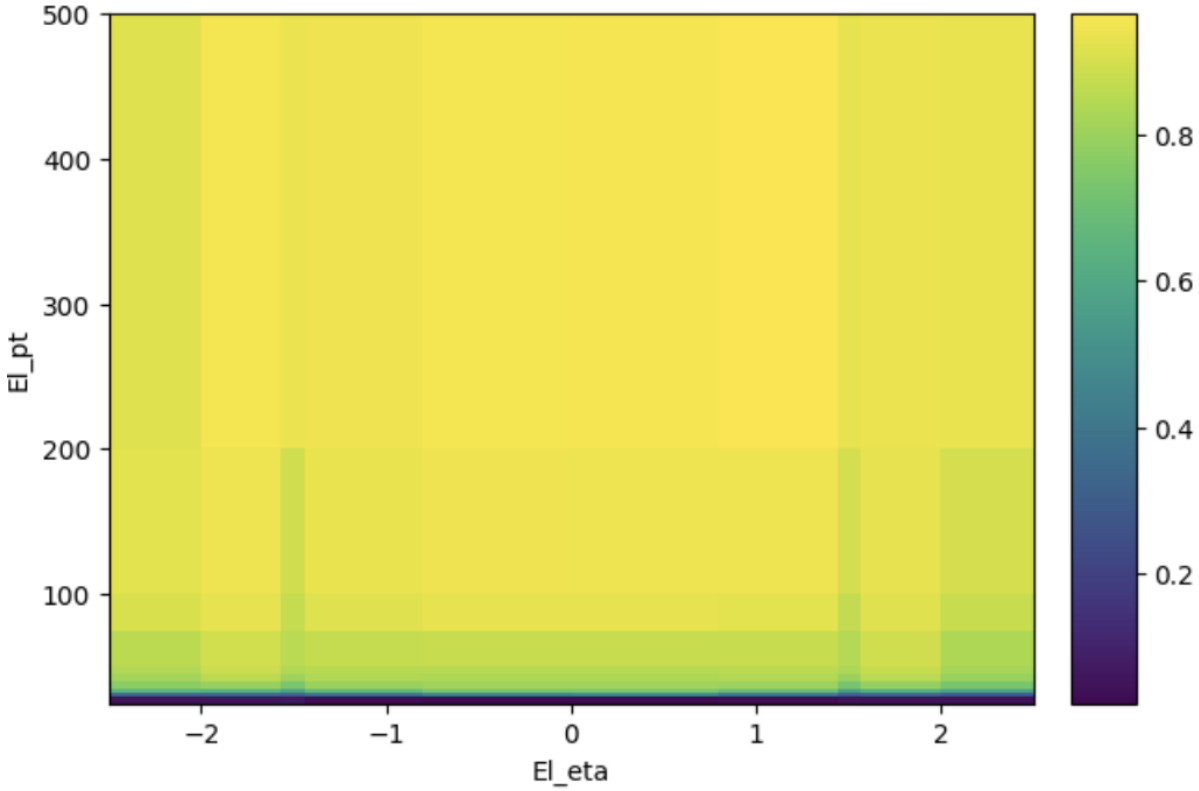


(b) 2022 Pre EE with extra parameters Standard Deviations

Figure 12: 2022 Pre EE with extra parameters



(a) 2023 Pre EE Efficiency in real data



(b) 2023 Pre EE Efficiency in simulated data

Figure 13: 2023 Pre EE Efficiency