The review here is divided into three portions separated by ++++ marks. (1) Comments from the perspective of time domain astronomy: AGN variability, DRW-type modeling, LSST cadences. This portion ends with an Overall Evaluation recommending a revised tone of the Abstract and Conclusions. (2) Comments from the perspective of the field of time series analysis where there is vast relevant experience in fields like signal processing and econometrics. These critiques apply to many AGN variability studies. (3) Minor textual matters.

++++++++++++++++++++++++++++

(1) Comments on the manuscript:

2.2.2 The section begins saying DHO models are "a more complex CARMA process might be a better model for describing quasars variability, especially for objects with weak oscillation features." This statement is a bit misleading: autoregressive models as simple as AR(2) commonly produce quasi-periodicities. See e.g. Chen & Li (ADS: 2005ChJAA...5..495C) and https://web.stat.tamu.edu/~suhasini/teaching673/chapter3.pdf. AGN (particularly blazar) astronomers who don't know this mathematical property of AR stochastic models often mistakenly conclude that a deterministic physical periodic process (e.g. orbiting supermassive black holes) must be present if a (quasi-)periodicity is seen in the lightcurve.

2.2.3 This section might be expanded for non-specialists. GP regression (a high-dimensional local regression) and CARMA models (a low-dimensional global regression) appear to be quite different approaches to modeling time series. The connection between the two is given in sec 7 of Forman-Mackey+ 2017, but I personally don't understand why this is true.

4.2 & 5.1.1 "As the goal of applying SRNN is to model the whole light curves, including those dates where there are no observations", the authors set missing timesteps to the mean value. I am concerned about this crucial element of the analysis in two respects:

(a) The choice of imputation with mean values is more simplistic than the ARMA- or NN-based techniques used by other researchers. Imputation (i.e. interpolation) of gaps in an evenly spaced time series is a well-known procedure. When stochastic autoregressive processes are involved, the imputed values are based on ARIMA modeling. This was performed by the Kepler team for their 4-year photometry of ~200K stars, and has many code implementations used in other applications (R/CRAN packages imputeTS, amelia, imputePSF, mtsdi and imputeTestbench). Imputation based on LSTM or GRU-AE neural networks are described by Li+ (DOI: 10.1016/j.neucom.2020.05.033) and Shu+ (10.23919/CCC52363.2021.9550206). Can the authors examine whether a more sophisticated imputation procedure may improve the fitting?

(b) The gaps in LSST WFD cadences are so great that any imputation technique is likely to fail. For example, the forecasting (similar to imputation) period in Yin+ 2021 (final figure; this is the paper providing many of the methods used by the authors) is applied to only ~10% of the duration of filled data. The problem is evident in the 4th panel of Fig 8 around mjd=2400-3200. The issue is not that the overall error (MAE) increases with gap size, but that short-lived events (100-200 day) occurring during gaps are entirely missed and the consequent model fits are

completely wrong.

The problem is not severe when the AGN variations are slower (>1 year) as in Figs 9-10. The problem also does not appear in the DHO model reconstruction in Fig 11, but I worry this is artificial: the authors first simulate, then fit, a global model where the quasi-periodic timescale is fixed. But in a real AGN, particularly when the quasi-periodicity arises from stochastic effects and/or loses phase (e.g. disk turbulence rather than a deterministic effect such as a radiant concentration at a fixed radius in the disk), this model will probably predict non-existent variations and fail badly.

This problem is not unique to the author's particular modeling approach, but will affect all LSST analysis of variable objects. The effect on LSST AGN identification, characterization and classification is not nearly as bad as it is for variable stars where the characteristic timescales of variations are typically shorter than the gap durations. Nonetheless, an improved discussion of the 'filling the gap' problem is needed. Perhaps the authors' method can provide some measure of when the fit is reliable (Figs 10-11) and when it is not (4th panel of Fig 4).

5.2.1 In Fig 12, is there a reason why the CARMA Metric (eqn 7) is never less than one? Is this somehow related to the desired fit metric \chi^2_\nu = 1.0 in least squares modeling which is the optimal variance per degree of freedom of (data-model) residuals? Or is this a coincidence?

5.2.2 (a) In Fig 13, I don't understand why, for both the LSST data alone and for the SRNN models, the output variability timescale is systematically lower than the input timescale for long input timescales. For example, if one were to input a \tau_in=5 yr with a LSST WFD cadence, I would expect both approaches to give reasonably accurate timescale recoveries. Instead, they often give \tau_out~2 years.

(b) Is it unfair to conclude from Fig 13 that the SRNN models are useless for estimation of \tau_in? A glance at the figure suggests that the colored curves are nearly flat, while the correct answer is a 45degree diagonal. The grey curves are closer to the diagonal and should be preferred. The stated improvement of SSRN models over LSST data in \tau recovery at long timescales is also consistent with the following (scientifically useless) interpretation: "SRNN always give \tau_out ~ 2 years irrespective of the AGN behavior for any LSST SFD cadence".

As outlined in item d) below, I recommend deleting Fig 14 and discussion relating to the structure function. If needed, replace with discussion of the ACF. Is the problem found here ("a positive correlation between SF∞ and τ") one of the problems discussed by Emmanoulopoulos+ 2010?

6.2 The comparison with previous work should include a discussion of Hu/Tak (ADS: 2020AJ....160..265H). They fit CAR(1) models in a state space representation to irregular cadenced *multiband* lightcurves; this is a multivariate generalization of Kelly+ 2009. While they do not simulate an LSST cadence directly, their sec 5.1 shows considerable improvement in obtaining decay timescales from the multiband model compared to univariate models for sparse, gapped cadences. In my opinion, the Hu/Tak approach is promising, and might be fruitfully

melded with the authors approach.

Overall evaluation: I find this study to be a sophisticated and capable analysis of LSST cadences on AGN variability. The issue is important: the effect of different LSST cadences on different scientific objectives is the subject of a Special Issue of the ApJSuppl now in preparation (https://community.lsst.org/t/apjs-focus-issue-dedicated-to-the-rubin-lsst-survey-strategy-optimization-process/5299). Recent studies relevant to this work are on arXiv (2105.12420, 2105.14889 accepted to MNRAS) but the ApJSuppl papers are not yet available.

The text can be improved, both with more discussion of broad time series issues (below) and by clarifying specific points in the study (above).

However, I recommend that the authors and Editors consider a major revision to the tone of the Abstract and Conclusions, giving a more gloomy view of the situation. The conclusion should be honest: Not much will be learned about AGN variability from the LSST WFD survey. There are 4 reasons for this recommendation:
(i) Item (4.2 & 5.1.1) above suggests that any attempt to 'fill the gaps' of LSST WFD lightcurves, with any of the examined cadence pattern, will miss AGN brightness variations with timescales shorter than ~1 year. This is a major lapse.
(ii) Item 5.2.2(b) above suggests that the elaborate parametric (CARMA) and machine-learning (SRNN) modeling efforts are nearly useless in estimating timescales of variation.
(iii) No single LSST cadence strategy appears much stronger than the others (Conclusions, final bullet). I would have hoped that 'cadence_drive' or 'rolling' cadences would do better than 'baseline', but this was not found by the authors.
(iv) Unmentioned in the manuscript, all model results will be weaker for the vast majority of LSST AGN with magnitudes 22-24 where measurement error is important. All simulations are for mag ~ 20 objects where the photometry is precise.

Thus the methods here are weak at 'characterizing' AGN variability in a reliable, quantitative fashion. I would guess that the methods would be more effective at 'identifying' and 'classifying' typical AGN from the sea of other cosmic species (variable stars, supernovae, transients, anomalies like Changing Look AGN) ... but that is not addressed by the work here.

My gloomy view is not based on a highly critical opinion of the authors' methodology. I suspect that any serious effort to characterize AGN variability from LSST WFD cadences will encounter similar problems, unless the Hu/Tak multiband approach proves effective.

++++++++++++++++++++

(2) I discuss four issues in time series analysis are inadequately treated in the study. These issues are inadequately treated here and in most AGN variability studies. Time domain astronomers, especially methodological experts like the authors, should become familiar with the texts below. To take advantage of these approaches, Python codes are inadequate and more extensive codes in Matlab and R are needed; see e.g. R's 'CRAN Task View: Time Series Analysis" and Matlab's 'Signal Processing Toolbox'.

- Chatfield & Xing 2019, The Analysis of Time Series: An Introduction with R (7th ed, elementary text, 8K citations)
- Box, Jenkins et al., 2015, Time Series Analysis: Forecasting and Control (5th ed, authoritative text, 56K citations)
- Hyndman & Athanasopoulos, Forecasting: Principles and Practice, 2018 (3rd ed, 4K citations. Available at otexts.org/fpp3, includes cookbook for CRAN package 'forecast')

a) Nonstationarity in AGN variability. The CARMA models defined in Table A1 show how random shocks and recent past values influence future values of the time series based on \tau or \alpha coefficients. The model thus assumes that perturbed values revert to a global mean; i.e. the time series is stationary (discussed in Kelly et al. 2014). But there is no reason to believe that AGN output should be stationary (why should accretion onto the black hole be constant on multiyear timescales?) and they often are not stationary (most dramatically, Changing State AGNs). Most commonly in time series analysis, nonstationarity is reduced by subtracting the narrowest median filter (differencing operator, the "I" in ARIMA), but removal of any fitted smooth local regression model (e.g. GP regression) can also work. Maximum likelihood estimation (MLE) for the broad CARFIMA family is available: Tsai & Chan (2005, doi:10.1111/j.1467-9868.2005.00522.x) with code in the CRAN package 'carfima'.

b) Model selection. When LSST data emerges, CARMA models will be fitted to the irregular light curves as in Figs 6, 8 & 10. The fitting procedure for CARMA is MLE, or a Bayesian variant if prior constraints on parameters are available. This is not mentioned in the text. In standard practice for regularly spaced time series (texts above), the choice between autoregressive models, such asARMA(1,0) (=DRW) and ARMA(2,1) (=DHO), is based on a penalized likelihood measure like the Akaike Information Criterion. Best-fit models are calculated for a range of (p,q) and, using the AIC, the balance between parsimonious vs. complex models is determined by the data not the scientist. Model selection thus allows astrophysical insight into the accretion process (e.g. presence or absence of a harmonic oscillator). The AGN community has had needless debates on the adequacy of simple vs. complex autoregressive models because quantitative model selection measures are not widely used.

c) Model validation. Irrespective whether DRW or DHO models are preferred for an observed lightcurve, goodness-of-fit tests will be needed to show that the best-fit model is adequate. (This is not needed for model-based simulated lightcurves like Fig 6/8/10). A simple Anderson-Darling goodness-of-fit test of the cumulative observed vs. model brightnesses is a reasonable tool, as well as more powerful residual diagnostics (e.g. Ljung-Box test for autocorrelation and augmented Dickey-Fuller test for stationarity). These are not obscure methods: LB and ADF tests have millions of Google hits.

d) Structure function. The SF is missing from all texts on time series analysis. The usual formulation contains the same information as the AutoCorrelation Function; higher-order SFs give more information for large datasets (e.g. plasma turbulence studies.) The advantage of the ACF is the breadth of methodology; e.g. the LB test. The value of the SF for irregular cadenced AGN studies has been justifiably criticized by Emmanoulopoulos+ (ADS: 2010MNRAS.404..931E).

++++++++++++++++++++++

(3) Minor comments:

4.1 The reference for stochastic recurrent neural networks (SRNNs) Fraccaro et al. is NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems 2016, 2207–2215. The seminal paper for the method is: Bayer & Osendorfer 2014 arxiv.org:1411.7610. Note that Goyal/Bengio+ (2017, DOI: 10.5555/3295222.3295416) propose a computational method for SRNN's that may be better than the loss function in eqn 6. But the LSST datasets may be so sparse so that improved computational efficiency is not needed.

Table 1 Add GRU and LSTM

Table A1 The top 4 sections define the statistical models while the bottom 3 sections are nonparametric transforms of the model. This difference might be noted.